

REMARKS

This Response is being filed in response to the outstanding Office Action, dated December 16, 2005, in connection with the above-identified application.

In this Response, Applicants amend claims 48 and 52, and cancel claims 49, 51 and 53-57. Applicants amend and cancel the claims solely to expedite prosecution in view of the Office Action and the telephone interview that was held on May 15, 2006. Applicants do not acquiesce in any of the Examiner's rejections. Applicants reserve the option to further prosecute the same or similar claims in the present or a subsequent application.

Further, silence with regard to Examiner's rejection of a dependent claim, when such claim after amendment depends from an independent claim that Applicant considers allowable for reasons provided herein, is not an acquiescence to such rejection of the dependent claim(s), but rather a recognition by Applicant that such previously lodged rejection is moot based on Applicant remarks and/or amendments relative to the independent claim (that Applicant considers allowable) from which the dependent claim(s) depends.

Upon entry of the Amendment, claims 48 and 52 are pending in the present application.

The amendment to claim 48 is supported throughout the application. No new matter has been added. In particular, the amendment substantively incorporates into claim 48 the limitations of canceled dependent claims 51, 53, 55 and 56. Without limitation, the Examiner's attention is called to pages 17-18 of the application concerning the correlation of greyscale intensities of pixels in images with respect to the amendment incorporating the substance of dependent claim 56 into claim 48.

Telephone Interview

Applicants thank the Examiner and the Supervisory Examiner for the courtesies extended during the telephone interview conducted on May 15, 2006 with Applicants' undersigned representative. During the interview, Applicants' representative and the Examiners discussed potential amending language in light of the Office Action. This Amendment contains amendments to the claims based upon that discussion and the prior Office Action.

Claim Rejections - 35 U.S.C. § 103(a)

Amended Claim 48

The Examiner rejected claim 48, prior to the amendment herein, under 35 U.S.C. § 103(a) as being unpatentable over Johnston et al. (US 6,791,531) in view of Maruno et al. (US 6,191,773). The Examiner asserted that Johnston disclosed all of the limitations of claim 48, except limitation (e), and that Maruno disclosed limitation (e).

Limitation (e) of claim 48, *as amended*, is: “emulating a use of a click from the mouse to provide an input signal to the computer program, by providing an input signal in response to the location of the feature in the video image being confined to a region defined by a radius for a defined period of time.”

Without waiving their position that Maruno does not disclose limitation (e) of claim 48 as previously presented, Applicants respectfully suggest that Maruno does not disclose this limitation (e) of claim 48 *as amended*, and claim 48 *as thus amended* therefore is not unpatentable over Johnston in view of Maruno.

The Examiner stated that Maruno disclosed limitation (e) prior to amendment at column 8, lines 30-50. That excerpt discloses that the “number of projected fingers, that is the shape of the hand can be accurately judged.” It goes on to teach that “[o]n the basis of the position or shape of the hand ... a virtual switch shown on the display screen can be selected.” (col. 8, lines 29-33) This excerpt does *not* state that the hand must be held still, or be confined to a region defined by a radius for a defined period of time, for the signal to be generated. On the contrary, the excerpt falls in a portion of Maruno describing an embodiment that it describes as including “a motion recognizing unit for recognizing the *shape and/or move* of the hand.” (col. 7, line 65 to col. 8, line 2) (emphasis added) Moreover, the beginning of the cited excerpt occurs in a paragraph which itself begins by stating that “the user instructs by hand *gesture*” (col. 8, lines 14-16) (emphasis added), and continues by stating that plural shape filters, composed of plural band pass filters differing in band, may be used so that “the *motion* of the user may be judged,” (col. 8, lines 38-40) (emphasis added). Finally, the cited excerpt goes on to state that a waveform may be generated reflecting “*undulations* of the hand” or “*undulations* of the finger.” (col. 8, lines 47-50) Thus, in the excerpt cited by the Examiner, Maruno discloses generating a

signal in response to the position or shape of the hand, or the motion of the hand or fingers, *rather than by reference to the image being confined to a region defined by a radius* as is required by limitation (e) of claim 48 as amended, *and it nowhere states that the hand must be kept within that region for a defined period of time to generate a signal.*

Moreover, even aside from the fact that Maruno does not disclose limitation (e) as amended, it would not be obvious to combine Maruno with Johnston to achieve the invention of claim 48 because Johnston actively teaches *away* from the technique for generating a “click” – a feature in the image being confined to a defined region – disclosed in the methods herein and claimed in *amended* claim 48 herein. In the “Summary of the Invention,” Johnston describes the “clicking function” as generating a signal “in response to a **change** in the control object image.” (Col. 6, ll. 40-45) (Emphasis added.) The Summary of Invention in Johnston does not suggest generating a “click” in response to a feature in the image being confined to a region defined by a radius rather than changing or moving an undefined and unlimited amount. Accordingly, even if Maruno taught limitation (e) as amended, which it does not, a user of skill in the art would not be motivated to combine Maruno with Johnston and in fact would be taught by Johnston *not* to do so.

There are further reasons why claim 48, as amended, should be allowed.

First of all, Applicants have now amended claim 48 to substantively incorporate the limitation of former claim 56:

“determining a subsequent location of the feature in a video image from the video camera at a subsequent given time, by correlating greyscale intensities of pixels in trial subimages of the video image at the subsequent given time, with greyscale intensities of pixels in a subimage including the chosen feature in the video image at the first time, and selecting the trial subimage of the video image at the subsequent given time which has the highest correlation to the subimage including the chosen feature in the video image at the first time.”

In the Office Action, in finding former claim 56 unpatentable, the Examiner asserted that the substance of this limitation was disclosed by Johnston at col. 7, lines 35-48. Respectfully, Applicants disagree. In the cited portion of Johnston, there is no reference to greyscale intensity,

or to correlating greyscale intensity. Indeed, in Johnston, the feature of interest is located in the camera field of view by selecting a threshold brightness value, and constructing an image consisting of the pixels with brightness values exceeding that absolute value. See col. 6, ll. 1-4 (“signal thresholds to distinguish object images having an intensity above a predetermined threshold intensity”); col. 7, ll. 23-25; col. 12, ll. 10-13, 19-22; col. 14, l. 65 to col. 15, l. 5; col. 21, ll. 3-5, 11-14. The only alternative Johnston suggests is inverting the process, and selecting instead “a negative or dark image.” Col. 15, ll. 6-7. *Johnston does not suggest the use of correlation techniques lacking brightness thresholds to locate features, or any correlation of greyscale intensities.*

Since Johnston does not teach or suggest the subject matter of former claim 56, amended claim 48, which now incorporates this limitation substantively, is allowable over Johnston in view of Maruno.

In addition, claim 48 has been amended to include the limitations of former claims 51, 53 and 55:

51: the system is a computer program

53: in the step of choosing, the feature associated with a system user includes at least a portion of one of the system user’s head or face.

55: the video images from the video camera are formed by reflection of ambient light from objects in the video camera field of view including reflection from the feature associated with the system user.

Applicants respectfully suggest that no prior art, or combination of prior art not requiring hindsight, teaches a system incorporating these features together with those previously set forth in claim 48.

Amended Claim 52

The Examiner rejected claim 52 as unpatentable over Johnston in view of Maruno in view of Dupouy (US 6,057,845). It will be appreciated that, in view of the fact that as shown

above *amended* claim 48 is not unpatentable over Johnston in view of Maruno, claim 52, which depends from *amended* claim 48, is not unpatentable over Johnston in view of Maruno in view of Dupouy. Accordingly, claim 52 should be allowed.

Summary

Claim 48, *as amended*, is allowable.

a) Maruno does not disclose generating an input signal in response to the location of a feature in a video image being confined to a region defined by a radius for a defined period of time.

b) In addition, Johnston does not disclose determining the location of the feature in the video image at a given time by correlating greyscale intensities of pixels in trial subimages of the video image at the given time, with greyscale intensities of pixels in a subimage including the chosen feature in the video image at a previous time, and selecting the trial subimage of the video image at the given time with the highest correlation to the subimage including the chosen feature in the video image at the previous time.

c) In addition, claim 48 has been amended to incorporate the limitations of former dependent claims 51, 53 and 55 that the system is a computer program, the feature associated with the user includes at least a portion of the system user's head or face, and the video images from the video camera are formed by reflection of ambient light from objects in the video camera field of view including reflection from the feature associated with the user. No prior art, or combination of prior art not requiring hindsight, teaches a system incorporating these features and those set forth above.

Claim 52 is allowable. It depends from allowable *amended* claim 48.

Previously-Submitted Supplemental Information Disclosure Statements

Applicants have previously submitted a Information Disclosure Statement on January 10, 2002, and Supplemental Information Disclosure Statements on January 25, 2002 and March 29, 2002. The Examiner considered the initial IDS in connection with the initial Office Action herein. Applicants called the two Supplemental Disclosure Statements to the attention of the Examiner on September 6, 2005 in connection with the filing of the Request for Continued

Examination (RCE). However, Applicants have not received any indication that the Examiner has considered the two Supplemental Information Disclosure Statements.

Copies of the Supplemental Information Disclosure Statements are submitted herewith. Applicants respectfully request both be considered. Their initial submission was timely.

CONCLUSION

In view of the foregoing amendments and remarks, Applicants consider the Response herein to be fully responsive to the referenced Office Action, and respectfully submit that the pending claims are in condition for allowance. Early and favorable reconsideration is therefore respectfully solicited. If there are any remaining issues or the Examiner believes that a telephone conversation with Applicants' attorney would be helpful in expediting the prosecution of this application, the Examiner is invited to call the undersigned at 617-832-1118. Should an extension of time be required, Applicants hereby petition for same and request that the extension fee and any other fee required for timely consideration of this application be charged to Deposit Account, **No. 06-1448**.

Respectfully submitted,

Date: May 16, 2006
Customer No: 25181
Patent Group
Foley Hoag, LLP
155 Seaport Blvd.
Boston, MA 02210-2600

/sdeutsch/
Stephen B. Deutsch, Reg. No. 46,663
Attorney for Applicants
Tel. No. (617) 832-1118
Fax. No. (617) 832-7000

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:

Gips and Betke

Serial No: 09/892,254

Filed: June 27, 2001

Title: *Automated Visual Tracking For
Computer Access*

Docket No.: BOK-002.01

Group Art Unit: 2173

Examiner: To Be Assigned

CERTIFICATE OF FIRST CLASS MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as First Class Mail, postage prepaid, in an envelope addressed to: Commissioner for Patents, Washington, D.C. 20231 on January 25, 2002.


Michael Phelan

Commissioner for Patents
Washington, D.C. 20231

**SUPPLEMENTAL INFORMATION
DISCLOSURE STATEMENT**

Sir:

Pursuant to 37 C.F.R. § § 1.56 and 1.97 (b)(3), submitted herewith on a Form PTO-1449 is a list of publications known to Applicants and their Attorney. A copy of each publication is also being submitted herewith. Applicants respectfully request that the Examiner consider the listed documents and indicate that they were considered by making appropriate notations on the attached Form 1449.

This submission does not represent that a search has been made or that no better art exists. Nor does it constitute an admission that each or all of the listed documents are material or constitute "prior art." If the Examiner applies any of the documents as prior art against any claim in the application and Applicants determine that the cited documents do not constitute

"prior art" under United States law, Applicants reserve the right to present to the Office the relevant facts and law regarding the appropriate status of such documents. Applicants further reserve the right to take appropriate action to establish the patentability of the disclosed invention over the listed documents should one or more of the documents be applied against the claims of the present application.

Under 37 C.F.R. § 1.97 (b)(3), this Information Disclosure Statement is being submitted before the mailing date of the first Office Action on the merits; therefore, no fees are believed to be due. However, the Commissioner is hereby authorized to charge any required fee to our Deposit Account, No. 06-1448.

Respectfully submitted,

FOLEY, HOAG & ELIOT, LLP

Dated: January 25, 2002

By: 

Robert W. Gauthier

Attorney for Applicants

Reg. No. 35,153

Customer No.: 25181

Patent Group

Foley, Hoag & Eliot LLP

One Post Office Square

Boston, MA 02109

Voice: (617) 832-1000

Facsimile: (617) 832-7000

20/489392.1

Computers Seeing People

Irfan A. Essa

■ AI researchers are interested in building intelligent machines that can interact with them as they interact with each other. Science fiction writers have given us these goals in the form of HAL in 2001: A Space Odyssey and Commander Data in Star Trek: The Next Generation. However, at present, our computers are deaf, dumb, and blind, almost unaware of the environment they are in and of the user who interacts with them. In this article, I present the current state of the art in machines that can see people, recognize them, determine their gaze, understand their facial expressions and hand gestures, and interpret their activities. I believe that by building machines with such abilities for perceiving, people will take us one step closer to building HAL and Commander Data.

Building machines that can see has been one of the most exciting and challenging research quests of the last 30 years. Much effort has been expended on "automatic deduction of structure of a possibly dynamic three-dimensional world from two-dimensional images" (Nalwa 1993). There has been considerable progress in the areas of object recognition, image understanding, and scene reconstruction from single and multiple images. This progress, coupled with the improvements in computational power, has prompted a new research focus of making machines that can see people; recognize them; and interpret their gestures, expressions, and actions. In this article, I present methods that give machines the ability to see people, understand their actions, and interact with them. I present the motivating factors behind this work, examples of how such computational methods are developed, and their applications.

The basic reason for providing machines the ability to see people really depends on the task we associate with a machine. An industrial vision system aimed at extracting defects on an assembly line need not know anything about people. Similarly, a computer used for e-mail and text writing need not see and perceive the user's gestures and expressions. However, if our interest is to build intelligent machines that

can work with us, support our needs, and be our helpers, then these machines should know more about who they are supporting and helping. If our computers are to do more than support our text-based needs such as writing papers, creating spreadsheets, and communicating by e-mail, perhaps taking on the role of being a personal assistant, then the ability to see a person is essential. Such an ability to perceive people is something that we take for granted in our everyday interactions with each other. This ability to perceive people and interact with them naturally is essential as we move toward building machines like HAL in 2001: A Space Odyssey and Commander Data in Star Trek: The Next Generation.

At present, our model of a machine, or more specifically of a computer, is something that is placed in the corner of the room. It is deaf, dumb, and blind and has no sense of the environment around it or of a person near it. We communicate with this computer using a coded sequence of tapings on a keyboard. Imagine a computer that knows you are near it, knows you are looking at it, and knows who you are and what you are trying to do. Such abilities in a computer are hard to imagine, unless it has an ability to perceive people. Research in speech recognition has made considerable progress toward perception of human speech (see Cole et al. [1995] for a survey). Commercial systems capable of word spotting and recognition of continuous speech are now available. Analysis of the video signal to perceive people has become a challenging and exciting research avenue for the field of computer vision, resulting in significant progress in the recent years.

To make machines that see people, the computer must first determine if someone is near it (where) and count how many people are in its field of view. The next step is to identify who the people are. After the computer has identified the people, it can interpret facial expression, hand gestures, and body language to

If our computers are to do more than support our text-based needs such as writing papers, creating spreadsheets, and communicating by e-mail, perhaps taking on the role of being a personal assistant, then the ability to see a person is essential.

determine what the people want or are doing in the scene and why. In the upcoming sections, I present the approaches to determine where, how many, who, what, and why with reference to people in a scene. The answer to each question is not possible independently, and each question depends on the other as dictated by the situation. Before getting into details, I briefly discuss the applications of such a technology.

Applications

Applications of computer vision methods aimed specifically at seeing people are many and encompass several different areas.

Effective human-computer interaction (HCI): Imagine computers that interact with you as we interact with each other, using speech and gestures. Such computers will know when you are looking at them, will be able to detect where you are pointing, and will interpret your gestures. These types of gestural interface are an integral part of a growing trend toward more human-centered interfaces in HCI research. Specific applications for this technology arise in areas where traditional interfaces such as the keyboard and mouse are not effective. Such techniques will allow us to move toward more noninvasive and unencumbered interfaces that allow for interactive visualization of multidimensional scientific data and user-centered direct interaction with virtual environments.

Smart and interactive environments: Machines that can see will aid us in developing smart rooms, rooms that know who is where and what they are doing. Such rooms can help monitor children, senior citizens, or physically challenged individuals and provide assistance and care as needed (Penland 1996). These types of system and the related interfaces could become a part of our daily activities.

Surveillance and security: A more traditional application of this work is surveillance and security. Face recognition has become quite a useful technology in the security industry, where access is allowed based on facial identity. Systems that automate searches of mug-shot databases to aid in criminal identification are being considered. Recently, work aimed at recognition of human actions promises great help for active surveillance applications.

Entertainment, education, and training: Two areas of recent rapid growth are education and entertainment. Computer vision methods for noninvasive tracking and interpretation of human activities can revolutionize various aspects of these areas too. An intelligent tutor that can see will be far more responsive to the

needs of its student if the tutor can judge by the actions and moods of the student whether he/she is confused, frustrated, or confident. Similarly, the development of complex environments for gaming and training will rely on the recognition and interpretation of actions and intentions of the user. A system that understands motions can aid in training for sports and teaching dance.

Video conferencing and model-based coding: Analysis and recognition of facial actions, gestures, and body language, especially with model representations of actions, would be useful for symbolic compression of video data. Vision-based methods for extracting spatiotemporal procedural information of hand gestures, body movements, and facial expressions will aid in the development of model-based video coding methods. With these methods, low-bit-rate videophones and model-based coding systems can be developed. The Moving Picture Experts Group (MPEG) (1999) community is already looking into these issues (see MPEG.org).

Digital libraries and video-image annotations: Automatic content-based annotation of images and video is an important application, especially as the amount of digital content grows at an exponential rate. Because a sizable portion of these data are about people, machines that can recognize people and their activities in images and video will play a significant role in the automatic annotation of these data.

Human augmentation and wearable computing: Systems that can interpret activities of the people in an environment could provide invaluable assistance to hearing-impaired or visually impaired individuals by translating the missing communication modality into a modality that they can directly understand. For example, a seeing computer might describe the body language of a conversational partner to a visually impaired individual through an earphone. The technology could also allow the impaired individual to communicate more effectively, for example, by translating sign language into spoken English (Stamer 1995). This form of intelligence driven by perceptual processing and aimed primarily at augmenting users, is becoming an important and challenging research area, especially as computers are taking on newer "roles," for example, wearable computing and affective computing (Wearable 1999; Picard 1998).

In the upcoming sections, I discuss the various aspects of research in computer vision that will play an essential role in the building of machines that can see people.

Is Someone There? Where? (Looking for People)

The first step toward building computers that are aware of people around them is to provide them with the ability to ask, Is someone there? Where? What is their location? Where are they looking? This is achieved by various methods, each varying in detail and function. The most common approaches include subtracting simple backgrounds, looking for specific color features, tracking motions, detecting changes, looking for faces, and tracking heads to determine a pose. I discuss these methods briefly here. First, I address methods for tracking whole bodies from imagery, then I present methods for tracking heads and determining head pose. Whole-body-tracking methods determine where people are, and head-tracking methods extract where people are looking.

People Tracking

The simplest methods for tracking people in a scene are based on image differencing. In these methods, the background image is acquired and stored before the person enters the scene. The person is then segmented in the image by subtracting each new incoming image with the stored background image, which extracts a silhouette of a moving person. A more general method for tracking people using this type of background subtraction requires modeling the scene as a set of distinct classes, including a background class and several classes that cover the person in the foreground.

The PFINDER system uses background and foreground classes to distinguish between the foreground silhouette and the fixed background (Wren et al. 1997). This operation provides the system with a background class while the person is modeled as a connected set of blobs in the foreground, each connected set defining a class. Each blob has spatial (x, y) and color (Y, U, V) properties. In each image of the scene, every pixel must belong to one of the classes. A representation of flesh colors is also encoded to aid in tracking hands and face. These blob features allow tracking of a person's hands and head from low-resolution imagery in real time. To aid in tracking, a low-level description of a model of a person—hands are on the sides, and the head is the highest point of the moving blobs—is also used. The approximate hand positions extracted in this way are used for static gesture recognition. Recent extensions to color-tracking methods include developing Gaussian mixture models of color space to extract flesh tones in the scene.

The basic limitation of the color-based track-

ing methods is the inherent limitations resulting from the use of color as a metric. Although skin color is a reliable feature for distinguishing between other parts of a person and the hands and face, it has serious problems when users wear skin-colored clothes or short-sleeved shirts and shorts. This limitation is addressed by combining various measurements, as discussed later.

A major advantage of such color-based foreground-background segmentation systems is that they can run in real time on simple desktop computers, allowing for easy development of simple systems for tracking people. Such color-based tracking systems have been demonstrated live during conferences and exhibitions (Darrell et al. 1998; Mase 1993a). However, a significant limitation still exists that current desktop computers barely allow for full-frame video capture (30 frames a second, 640 x 480 pixel resolution) in real time. Most color-tracking systems run at 10 frames a second on a 320 x 240 image and are sufficient for tracking people that don't move too fast.

Another limitation of using a single-camera system running a color-based blob tracker is that it requires a well-calibrated three-dimensional (3D) environment for 3D tracking of the user. This is addressed by running the same blob tracker on two different cameras and extracting positions of the person in 3D using image correspondences, triangulation, and camera-to-camera calibration. A wide-baseline stereo camera system can be used to self-calibrate such a scene, and then stereo matching can be used to track a person in real time. However, it should be obvious that color-tracking methods cannot directly be extended to track multiple people.

Reliable tracking of multiple people is achieved by implementing simple background subtraction techniques in a well-constrained and calibrated closed-world environment. In the KIDSRoom environment (Bobick et al. 1997), a complete domain of the scene being observed is defined, and then silhouettes are tracked over time. With simple metrics of velocity, occlusions are resolved. The domain and the storyboard of an interactive entertainment space are used to determine and control the activities of the participants (users) to aid in tracking.

In addition to the color-based methods, several other methods have been proposed. These methods use a detailed a priori structure of the person being tracked. Similar to the methods discussed earlier, these methods also extract image features—silhouettes, color, and edges—from a scene to aid in tracking people.

The first step toward building computers that are aware of people around them is to provide them with the ability to ask, Is someone there? Where? What is their location? Where are they looking?

Baumberg and Hogg (1994) present methods for using simple models and active contour representations to locate and track people. Bregler and Malik (1998) present a feature-based tracking method coupled with a kinematic model of a walking person to track people. Gavrilu and Davis (1996) present a method for tracking people from multiple views. These are more robust tracking techniques compared to the simple color-based tracking methods. In addition, these model-based methods are also more accurate at characterizing the motions. However, these methods are also more computationally complex and require special hardware to achieve real-time performance.

Finding Faces

With the methods described earlier, people are located by simply looking for specific colors or detecting a change in an image. There is no real notion of a person, except when defined a priori to aid in tracking.

A completely different type of method aimed at locating people uses an a priori model of a face and its features to search for a face over the whole image. These methods use features associated with facial shape to determine the number of faces in a scene (Boluja and Kanade 1998a, 1998b; Colmenarez and Huang 1997; Lueng, Burl, and Perona 1995; Moghadam and Pentland 1995; Turk and Pentland 1991). The techniques that are used in these methods are similar to the ones used in face-recognition methods and are discussed later.

These methods are not yet fast enough for real-time tracking and are presented mostly as a way of extracting faces from static and complex scenes. A real benefit of these systems is that most of these methods are reliable for locating multiple people in a scene. The increase in available computation power will allow for real-time application of these methods. These methods can be combined with the color-tracking or motion-change-detection algorithms to reduce the search space, as discussed later. These systems might serve as a precursor to the face-recognition system that answers the question, Who is in the scene?

Head-Pose Tracking

Determining where a person is and where a person is looking is extremely important for development of systems that are aware of people and are able to recognize the person's face and expressions. Most techniques for expression tracking and face recognition work reliably only for small head motions. This limitation reduces the applicability of these methods, and consequently, head tracking has become

an increasingly important research topic.

Head tracking can be achieved by observing a set of features on a face or warping a template of a face to match the transformations of the face as it moves. All the problems inherent in head tracking and pose determination are the same as in determining the orientation of an object for object recognition. Methods that attempt to extract complete 3D structure of the face from visible features to methods that match image templates with affine transformations have been developed. Azarbayejani et al. (1993) present a recursive estimation method for extracting structure and motion of a head by tracking small facial features such as the corners of the eyes or mouth. However, because of its dependence on feature tracking, its applicability is limited to sequences in which the same points were visible over most of the image sequence.

Black and Yacoob (1995) have developed a regularized optical-flow method that uses an 8-parameter 2D affine model of flow that yields good results for pose estimation. However, the use of a planelike 2D model limits accurate tracking to medium-size head motions, and the method can fail for very large head rotations.

Robust head tracking requires a technique that can be characterized as motion regularization or flow regularization (Essa et al. 1996). In this technique, flow between two frames is computed, and the rigid motion of the 3D-head model that best accounts for this computed flow is used as an estimate of head motion. The results of this model-based tracking are shown in figure 1, which shows five frames from a long sequence of a person moving his head.

Combining People Tracking with Face Finding

Robust methods for tracking multiple people using multiple cameras are also being developed. These methods rely on combining methods for color- and silhouette-based tracking with face-detection methods. Stillman, Tanawongsuwan, and Essa (1999) present a robust real-time method for tracking multiple people with multiple cameras. In this method, both static cameras and pan-tilt-zoom (PTZ) cameras are used to extract visual attention in the environment. The PTZ camera system uses face recognition (described in the next section) to register people in the scene and lock on to these individuals. A commercially available face-recognition system (Visionics 1997) that runs in near real time on a PENTIUM PC is used for face tracking and identification. This commercial system uses the video signal from the

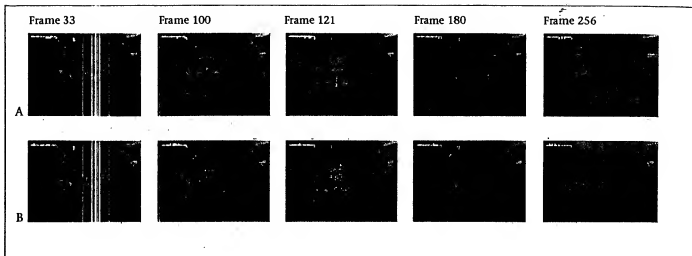


Figure 1. Results of Tracking a Sequence with an Ellipsoidal Model.

A. Original image sequence (300 frames). B. Tracking using three-dimensional ellipsoidal model.

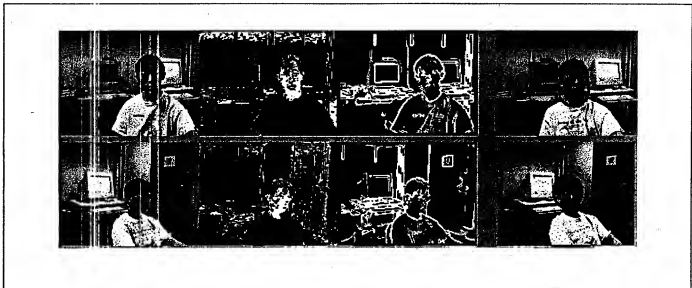


Figure 2. A System Tracking and Following a User's Face.

A combination of color segmentation, movement tracking, and shape information is used for robust tracking of a face.

PTZ cameras to find a face and adjusts the visual foveation process of the PTZ camera.

The static camera system provides a global view of the environment and is used to readjust tracking when the PTZ cameras lose their targets. The system works well even when people occlude one another. The underlying visual processes rely on color segmentation using blob tracking, movement tracking, and shape information to locate target candidates. Color-indexing and motion-tracking modules help register these candidates with the system, allowing for robust tracking. Results of this system are shown in figure 2 for tracking a face

using a single camera. The multiple-camera, multiple-people tracking system is described in figures 3 and 4. A distinctive advantage of this type of foveation mechanism is that in addition to a good estimate of the location of the person and his/her face, the system acquires a higher-resolution image of the face that can help with recognition or expression tracking.

Darrell et al. (1998) also present a method that combines face tracking (Rowley, Baluja, and Kanade 1998a) with color tracking to set up a multimodal system for tracking and identifying people.

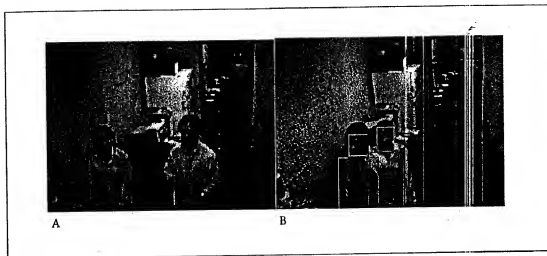


Figure 3. Results of a Multiple-People, Multiple-Camera Tracking.

A. Two people entering a scene are tracked as they move around. B. The two people occlude each other.

Who Is It? (Recognizing People)

Over the past 30 years, extensive research has been conducted by psychophysicists, psychologists, neuroscientists, and engineers on various aspects of face recognition by humans and machines (see Bruce [1988] and Ellis et al. [1986] for review of work on human perception of faces). The earliest work on machine recognition of faces appeared in the mid-1970s, when typical pattern-classification techniques were used to measure and compare facial-feature attributes for recognition (Kanade 1977). Not much work appeared in this area until the 1990s when the availability of increased computational power, coupled with a commercial demand of face-recognition systems, made the problem computationally viable and commercially exciting.

At present, face recognition is perhaps the most widely studied topic in the vision community. It has the distinct privilege of being the first application of computer vision to be commercialized that is not related to industrial machine vision. At last count, there were 19 commercial ventures attempting to bring face-recognition applications to the public (Face 1999).

The last few years of increased activity have seen progress in locating and segmenting a face in a complex scene; extracting features such as eyes, mouth, and nose; and recognizing occlusions and changes in facial features with orientation, pose, and scale variability. It should be noted that all these problems are standard problems also addressed in the traditional computer vision goal of object recognition and are now being applied to the newer

domain of faces. The face-recognition domain, because of its inherent applications, has resulted in significant advances in the design of statistical and neural network-based classifiers. Because of the existence of a large body of literature on a machine-vision method for face recognition, my exposition of this area is brief. Interested readers are encouraged to peruse survey publications by Chellappa, Wilson, and Sirohey (1995) and Samal and Iyengar (1992).

Pattern-Recognition Methods for Face Recognition

As stated earlier, face-recognition methods have resulted in significant developments in various pattern-recognition methods. Recently, a need for a suitable representation for detection and recognition of faces from images has generated renewed interest in Karhunen-Loeve expansion methods (Kirby and Sirovich 1990; Sirovich and Kirby 1987). Karhunen-Loeve expansion methods, also known as principal component analysis (PCA) methods, are widely used in the pattern-recognition area.

A PCA-based method called eigenfaces (Moghaddam and Pentland 1995; Pentland, Moghaddam, and Stamer 1994; Turk and Pentland 1991) for face recognition has shown very high recognition accuracy (around 95 percent) using databases of more than 7500 face images of about 3000 people. In this method, faces are aligned with each other and treated as high-dimensional pixel vectors from which eigenvectors (called *eigenfaces*) and eigenvalues are computed. These eigenvectors represent the principal components; therefore, the eigenvalue decomposition method allows for representing the probe face by a small number (some-

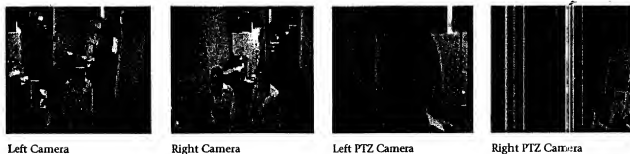


Figure 4. Views from Two Static Cameras (Left and Right) Showing the Result from the Person-Detection System. Also shown are views from the two pan-tilt-zoom (PTZ) cameras placed in the front of the room. A triangulation process is used to decide the scale factor (that is, the distance from a person to the PTZ camera).

times 100) of expansion coefficients, which are then used in recognition. The alignment of all the faces is done automatically by using a similar representation for facial features (eyes, nose, and mouth). Several extensions to these methods have recently been proposed (Etemad and Chellappa 1994; Swets and Weng 1996).

Another popular method relies on collapsing the variances in facial images to extract face descriptors called *image graphs*. In these graphs, facial features are described as a set of wavelet components. Image graphs are extracted by generating a *bunch graph*, which is constructed from a small set of sample image graphs. Comparison of this image graph between images yields recognition of facial images (Kruger, Potzsch, and von der Malsburg 1997; Wiskot et al. 1997). This work extends the work of Manjunath, Chellappa, and von der Malsburg (1992) that uses Gabor wavelet decomposition and that of Landes et al. (1993) that uses dynamic link architecture (DLA). Impressive results for recognition of faces from different viewpoints are reported.

In addition to classical pattern-recognition methods, much work exists on applications of neural networks for face recognition. Rowley, Baluja, and Kanade (1998a, 1998b) present good results for face detection using retinally connected neural nets that examine small windows of an image and decide whether each window contains a face. They use multiple neural nets and have shown reliable results with large variations in pose. Brunelli and Poggio (1993) present a different method using a HYPERBF network for recognition of a face.

Because of the large body of work on face recognition in recent years, it is almost impos-

sible to cover all the significant developments. However, it is important to observe that each system claims good results, and the authors freely discuss the strengths and weaknesses of each method. Until recently, there was no definitive way of comparing these results, which led to the Face Recognition Technology Program (FERET) evaluation sponsored by the United States Department of Defense. Phillips et al. (1998, 1997) present the methodology and the results of these tests. The FERET Program provides a methodology for reliable testing of different face-recognition systems over a large database (14,126 images of 1,199 people) collected independently. These tests are very successful in evaluating the state of the art in face-recognition methodologies and measure algorithmic performance over large databases. These tests rated the systems from the Massachusetts Institute of Technology (Moghaddam and Pentland 1995; Pentland, Moghaddam, and Starner (1994), the University of Maryland (Etemad and Chellappa 1996; Manjunath, Chellappa, and von der Malsburg 1992), the University of Southern California (Kruger, Potzsch, and von der Malsburg 1997; Wiskot et al. 1997), and Michigan State University (Swets and Weng 1996) as very proficient in recognizing faces.

Does that mean that face recognition is a solved problem? The evidence supports this to be true for face recognition in limited domains and applications with full frontal faces.

Under constrained environments with full-frontal faces, there is every reason to expect these face-recognition systems to perform reliably. However, much research is still needed to resolve face recognition in unconstrained environments with variations in lighting, orientations, and changes in facial features.

What Do They Want or What Are They Doing? (Gesture, Expression, and Activity Recognition)

Now, I present the methods for asking questions of what is happening in an environment. I start with a discussion about recognizing facial expressions, then explore gesture recognition and interpretation of human activity.

Facial Expression Recognition

The psychology community has a large body of work on face perception and facial analysis. Perhaps the most important work in this area is the effort led by Ekman, and Friesen (1978), who produced a system for describing all visually distinguishable facial movements called the facial action coding system (FACS). In this system, each expression can be represented in terms of action units. It is believed that automatic recognition of facial expressions from images can be achieved by categorizing a set of predetermined facial motions, such as with FACS, rather than determining the motion of each facial point independently.

Yacoub and Davis (1994), Black and Yacoub (1995), and Mase (1993b) use the FACS representation for recognition of facial expressions. Yacoub and Davis extend the work of Mase by detecting motion in six predefined and hand-initialized rectangular regions on a face and then use simplifications of the FACS rules for the six universal expressions for recognition. The motion in these rectangular regions from the last several frames is correlated to the FACS rules for recognition. Black and Yacoub extend this method further by using local parameterized models of image motion to handle large-scale head motions. These methods show about 89-percent accuracy in correctly recognizing expressions over their database of 105 expressions. They have also shown remarkable success at recognizing expressions from real video of people in television talk shows. These results are impressive considering the complexity of the FACS model and the difficulty in measuring facial motion within small-windowed regions of the face.

It has been argued that one of the main difficulties these researchers have encountered is the complexity of describing human facial movement using FACS. These limitations of FACS as a representation of facial motion for automatic recognition have recently generated a lot of discussion. National Science Foundation (NSF) workshops and the resulting reports on facial expressions discuss this issue

in detail (Pelachaud, Badler, and Viard 1994; Ekman et al. 1993).

Essa and Pentland (1995, 1994) and Essa, Darrell, and Pentland (1994) undertake detailed experiments for measuring facial motion and report that it is important to move away from a static, dissect-every-change analysis of expressions. This extension toward a whole-face analysis of facial dynamics in motion sequences is even more significant for machine perception of facial motion. They have analyzed video data of facial expressions and then probabilistically characterized the facial muscle activation associated with each expression. This characterization is achieved using a detailed, physically based dynamic model of the skin and muscles coupled with optimal estimates of optical flow in a feedback-controlled framework. A second, simpler representation that encodes the motion and velocity in the image plane is also extracted. There are 2D spatiotemporal templates to represent facial expressions.

This detailed analysis of video data yields two representations of facial motion that are then used to recognize facial expressions in two different ways. These extracted representations are graphically shown in figure 5. Both of these methods for recognition of facial expressions result in 98-percent accuracy over 52 image sequences. However, these results are preliminary, and comparison with other techniques is not possible without using all the proposed methods for facial-expression recognition on the same test set. A FERET type of initiative would be beneficial to this type of research.

One of the major problems with these facial-expression techniques is that they do not run in real time or even at interactive rates. At present, the method that uses a dynamic physics-based model of the face is computation intensive. On a Silicon Graphics INDY R5000 180-megahertz machine, each frame takes about 15 seconds. The method that uses the 2D spatiotemporal templates is much more efficient and runs about 5 frames a second (fps) (that is, 1 second of 160- x 120-resolution video at 15 fps would take 3 seconds after digitization). Using specialized hardware and multiple-processor Pentiums could aid in such computations.

Essa, Darrell, and Pentland (1994) and Darrell, Essa, and Pentland (1996) present a method for facial tracking and interactive animation of faces that runs in real time. The basic idea for this method is to do a fine-grained analysis of a subject's expression and then store the spatiotemporal representation

At present, face recognition is perhaps the most widely studied topic in the vision community. It has the distinct privilege of being the first application of computer vision to be commercialized that is not related to industrial machine vision.

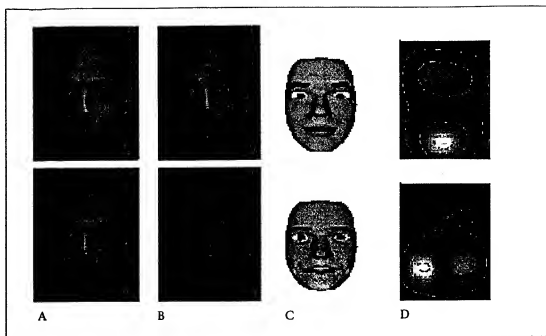


Figure 5. Determining Expressions from Video Sequences.

A. Neutral: The surprise expression showing in the top row, and the smile-happiness expression showing in the bottom row. B. Expression: The model used for analysis synthesis. C. Model. D. The motion energy: Peak muscle actuation and motion energy are used for recognition of expressions (Essa and Pentland 1997).

of this expression on a generic model of a face. Then simple visual measurements can be used to establish the relationship between an image and the dynamic motion parameters of the model. These simple visual measurements could be appearance and view based, feature based, blob based, or even motion based. These measurements are coupled with the parameters of the physics-based model using an interpolation process, resulting in a real-time, passive (that is, the observations drive the model) facial tracking and animation system (figure 6). In addition to tracking expressions using this method, hidden Markov models (HMMs) could be used for recognition of expressions based on a similar set of visual measurements.

It is important to note here that the previous methods are aimed at recognition of facial expressions. Because there is a known relationship between facial expression and human emotions (see Ellis et al. [1986], Bruce [1988], and Ekman and Friesen [1969]), it is foreseeable that such techniques can be used to recognize human emotions. Although the possibilities of developing such systems are both exciting and challenging (Picard 1998) and raise many intriguing social implications, not much work to date has been attempted to build and evaluate such a system. Building machines that can recognize emotions and

read lips is an actual goal of the work on the recognition of facial expressions.

Gesture Recognition

There are many facets to modeling, tracking, and recognizing human gesture and body motion. For example, gestures can be made by hands, faces, or the entire body; have strong spatial and temporal characteristics; can be person or culture specific; can be tied to a linguistic basis or spoken conversations; or can be meaningful in their own right. For this reason, research in several domains (vision, AI, linguistics, biomechanics, and robotics) is relevant for automatic understanding of gestures.

Many researchers in the vision community have attempted automatic gesture recognition and body tracking from video (Darrell, Essa, and Pentland 1996; Baumberg and Hogg 1994; Kakidiaris, Matas, and Bajcsy 1994; Rehg and Kanade 1994). In these efforts, pattern-recognition methods are applied to extract spatiotemporal codings from image streams for recognition. Learning algorithms have also been used for interpretation of gestures (Stamer, Weaver, and Pentland 1998; Yamato, Ohya, and Ishii 1994). This area of research has been furthered by successful attempts at using appearance-based or view-based methods for tracking and recognizing human motion (Black and Jepson 1996; Moghaddam and Pentland 1995).

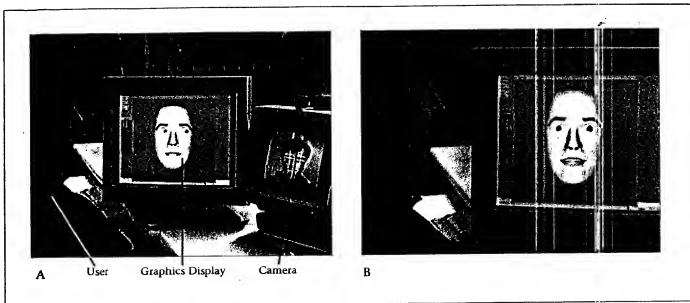


Figure 6. Real-Time Tracking of Facial Movements.

A. Complete system tracking eyes, mouth, eyebrows. B. Mimicking a smile expression.

The importance of time in the analysis and recognition of hand and body movements has led to the use of HMMs for recognition after training on views of the model. For example, Yamato, Ohya, and Ishii (1994) studied the recognition of tennis strokes by training on time-sequential images of six different tennis shots. Starner, Weaver, and Pentland (1998) use HMMs for recognition of American Sign Language. Bobick et al. (1997) and Bobick and Wilson (1995) have shown a unique way of representing gesture that captures both the repeatability and variability of gestures in a training set of example trajectories of gesture states.

Gesture understanding requires interpretation of the spatiotemporal patterns extracted from video with the constraints imposed by the dynamic representation of human action and the linguistic context, if any, of such an action. To achieve such an understanding of human gestures, we need to develop a theory of human action that has an inherent computational value. Essa and Pentland (1995) present a similar idea that relies on a computational value for interpretation of facial expressions. In this approach, a reduced dimensional representation of facial action is developed by a causal reconstruction of how the scene was produced. This representation is achieved by coding facial movements from video in terms of muscle contractions and using an analysis-synthesis framework. A similar attempt at a preliminary extension of this method (framework) for whole-body actions

by using a kinematic model of a human figure is presented by Brand and Essa (1995). Lakoff and Johnson's (1980) theory on metaphors for actions is used for empirically defining high-level human actions from low-level kinematic motions.

Tracking three-dimensional human movements from video is far from a trivial problem because tracking generally is an underconstrained problem, the data are noisy, and the measurements include several levels of nonlinearity. Adding a layer of constraints imposed by the dynamic representation of human action and the linguistic context of the action should help with analysis and interpretation.

If the interest is in recognizing and representing higher-level human actions, we can gain insight from research on how humans express themselves and how they move. Unlike the machine-vision community, the linguistics community has been studying the communicative aspects of gestures for many years (Kendon 1974; Ekman and Friesen 1969; Efron 1941). Some recent work is aimed at understanding gestures in the context of communication, especially speech (McNeill 1992; Krauss, Morrel-Samuels, and Colasante 1991; Cassell and McNeill 1990). We believe that this work provides us with at least a preliminary understanding of communication through gestures and should provide rules to help with the interpretation of gestures.

For more detailed analysis of human movement, we can rely on the biomechanics literature that provides motion-capture data, force-

plate data, and muscle-activation records. These data tell us how and why humans move in the ways they do. These data can be used to tune control algorithms for human motion and provide additional constraints on the candidate descriptions for a motion sequence.

In computer animation, researchers have explored (Hodgins 1998) the use of dynamic simulation as a technique for generating human motion for computer animation and virtual environments. These dynamic motion generators for human action can be extended to provide us with both a higher-level representation (behavior or activity level) and a lower-level description in space and time (joint angles, positions, and so on) for additional behaviors that are appropriate for any application domain. Additional behaviors such as sitting, walking, pointing, dancing, and gesturing will force us to address stylistic issues. Studying the use of this type of generated motion with an appearance- and motion-based extraction of events from video will yield interesting results.

Activity Recognition

As stated previously, computer vision is a critical technology for creating systems that can interact naturally and intelligently with people. In addition to finding, tracking, and recognizing people, we can use computer vision techniques to recognize human activities in an environment (Seitz and Dyer 1997; Bobick 1996; Polana and Nelson 1993). Such recognition of human activities requires the study of the dynamic relationship between human motion and objects in the scene. Additionally, to address the issue of recognition of human actions and activities, it seems essential to develop an adaptive approach that uses context as a means of deciding the most appropriate representation that will be used for recognition.

It seems apparent that understanding the dynamics of human motion is fundamental to solving action-recognition problems (see Cedras and Shah [1995] for a review). A common thread in much of the recent work in action recognition has been the use of HMMs as a means of modeling complex actions. Lately, there have been several contributions in the literature that offer new frameworks for activity recognition. Specifically, Bregler (1997) evaluates motion at graduated levels of abstraction by using a four-level decomposition framework that learns and recognizes human dynamics in video sequences. Although Bregler's method focuses on complex human motions, such as walking, Oliver, Pentland, and Berard (1997) present a system designed to

assess interactions between people using Bayesian approaches. Bobick (1996) also presents several approaches to the machine perception of motion and discusses the role and levels of knowledge in each. The framework proposed by Buxton and Gong (1995) uses Bayesian networks for surveillance activities in well-defined and constrained scenarios.

Context management plays a critical role in this process by supplying, maintaining, and discovering information about the relationships between people and objects. Objects provide clues about which human motions to anticipate, making them powerful tools for discriminating between actions and activities. Building a formal context model for people and their surroundings provides an architecture where acquired visual data can be warehoused, analyzed, and shared effectively.

To address this issue, Moore, Essa, and Hayes (1999) are developing an object-oriented approach called OBJECTSPACES to encapsulate context into scene objects. Instead of making static assumptions about the contents of an image sequence, they attach regions in an image of a scene to virtual objects. A scene object is derived as an instance of a class type. For example, if the scene is an office environment, classes would include desk, bookcase, and keyboard. All scene objects are provided with a priori information about the image regions they represent. By monitoring these regions, objects can develop an awareness of their features and can detect when their state changes. For example, if a person moves a book resting on a table by a few inches, the book object can determine that it has been moved and attempt to recalculate its new position. Additionally, each object understands complex actions that are indigenous to its class. For example, the book object stores a profile of two motion gestures—(1) page forward and (2) page backward—that it can identify by observing how humans interact with it. By evaluating these two actions over time, the book can decide if someone is quickly browsing through its pages or carefully studying every word. Tracking and motion analysis, which takes place in the extraction layer, is shared among objects representing scene articles and people. The scene's objects report their observations to a scene-level object, or *scene layer*, that catalogs all the activities taking place. This layer searches for correlations between object interactions to classify particular activities or identify certain human behaviors.

To test these representations, experiments in natural environments, where people interact with their surroundings, are recorded. The first

The psychology community has a large body of work on face perception and facial analysis. Perhaps the most important work in this area is the effort led by Ekman, and Friesen (1978), who produced a system for describing all visually distinguishable facial movements called the facial action coding system (FACS).

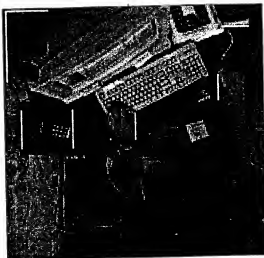


Figure 7. Recognition of Activities around a Workstation.

Using the ObjectSpace representation, objects are marked and their context established. Hands are tracked using color features, and activities are recognized.

experiment is based in a real office environment equipped with typical objects and appliances. Objects in the scene and the action associations are predetermined. Testing on a 5-minute video sequence of a user interacting with different objects in the office scene, the system detected 92 percent of the events correctly (figure 7).

This type of research effort, which is aimed at the recognition of human activity, has many practical applications where passive, nonintrusive action recognition is desired, such as video surveillance and activity annotation. Moreover, work conducted in this area advances computer awareness, which is an essential step toward the building of intelligent machines that can perceive and communicate with us naturally.

Conclusions

In recent years, there has been significant progress in the building of machines that are aware of the users who interact with them. The increase in computational power, combined with the multimedia capabilities of computers, has had a strong impact on the growth of this research area. Development of computer vision systems that are more than simple prototypes, and have applications beyond industrial vision, has been a boon for computer vision. In fact, it has resulted in a significant growth in computer vision research in recent years, mostly supported by industrial funding in addition

to the more traditional government funding. It is important to also note that current multimedia computers are also making it very easy to develop vision-based systems for interaction. Applications of this type of computer vision research are many and far reaching.

On a technical level, this domain of computer vision research has revived the concepts of pattern recognition for interpretation of a scene. Face-recognition methods are a perfect example of this revival and are aimed at the static interpretation of an image. In addition, some of the recent work in gesture and action recognition requires a study of dynamic signal and symbol interpretation. Research in several domains (vision, AI, linguistics, biomechanics, and robotics) is essential and relevant if we are interested in building machines that can see us.

These are indeed exciting and challenging problems and exciting and challenging times for research in computer vision. The day is not far away when our desktop computers will be able to see when we are looking at them. The next step toward building HAL and Commander Data is to make these systems robust to varying conditions and responsive to us in real time.

Acknowledgments

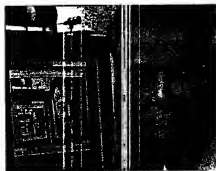
Special thanks to the students who have been involved in these projects at Georgia Tech, specifically the students in the Computational Perception Laboratory: G. Brostow, D. Moore, W. Rungtarityotin, A. Schoedel, D. Steedly, S. Stillman, A. Stoytchev, K. Sukel, and R. Tanawongsuwan. Also, thanks to Sandy Pentland (MIT Media Lab) under whose guidance some of the earlier work was undertaken, and thanks to our collaborators M. Brand (Mitsubishi Electric Research Labs), S. Basu (MIT), T. Darrell (Interval), and A. Ram (Georgia Tech).

References

- Azarbayejani, A.; Horowitz, B.; and Pentland, A. 1993. Recursive Estimation of Structure and Motion Using the Relative Orientation Constraint. In Proceedings of the Computer Vision and Pattern Recognition Conference, 15-17 June, New York.
- Azarbayejani, A.; Stamer, T.; Horowitz, B.; and Pentland, A. P. 1993. Visually Controlled Graphics. *IEEE Transactions on Pattern Analysis* 15(6): 602-605.
- Baumberg, A., and Hogg, D. 1994. An Efficient Method for Contour Tracking Using Active Shape Models, TR-94.11, School of Computer Studies, University of Leeds.
- Black, M. J., and Jepson, A. D. 1996. EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. *International Journal of Computer Vision* 26(1): 3-84.
- Black, M. J., and Yacoob, Y. 1995. Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using

- Local Parametric Model of Image Motion. In *Proceedings of the International Conference on Computer Vision*, 374-381. Washington, D.C.: IEEE Computer Society.
- Bobick, A. F. 1996. *Computers Seeing Action*. Technical Report, 394, Media Laboratory, Perceptual Computing Section, Massachusetts Institute of Technology.
- Bobick, A. F., and Wilson, A. D. 1995. A State-Based Technique for the Summarization and Recognition of Gesture. In *Proceedings of the International Conference on Computer Vision*. Washington, D.C.: IEEE Computer Society.
- Bobick, A., Intille, S.; Davis, J.; Baird, E.; Pinhanec, C.; Campbell, I.; Ivanov, Y.; Schutte, A.; and Wilson, A. 1997. *The kidsroom: A Perceptually Based Interactive and Immersive Story Environment*. Technical Report, 398, Media Laboratory, Perceptual Computing Section, Massachusetts Institute of Technology.
- Brand, M., and Essa, I. 1995. Causal Analysis for Visual Gesture Understanding. Paper presented at the AAAI Fall Symposium on Computational Models for Integrating Language and Vision, 10-12 November, Cambridge, Massachusetts.
- Bregler, C. 1997. Learning and Recognizing Human Dynamics in Video Sequences. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 568-574. Washington, D.C.: IEEE Computer Society.
- Bregler, C., and Malik, J. 1998. Tracking People with Twists and Exponential Maps. In *Proceedings of Computer Vision and Pattern Recognition*, 8-15. Washington, D.C.: IEEE Computer Society.
- Bruce, V. 1988. *Recognizing Faces*. Mahwah, N.J.: Lawrence Erlbaum.
- Brunelli, R., and Poggio, T. 1993. Face Recognition: Features versus Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(10): 1042-1052.
- Buxton, H., and Gong, S. 1995. Advanced Visual Surveillance Using Bayesian Networks. In *Proceedings of the IEEE Workshop on Context-Based Vision*. Washington, D.C.: IEEE Computer Society.
- Cassell, J., and McNeill, D. 1990. Gesture and Ground. In *Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society*, eds. K. Hall, J.-P. Keonig, M. Meachman, S. Reinman, and L. Sutton, 57-68. Berkeley, Calif.: Berkeley Linguistics Society.
- Cedras, C., and Shah, M. 1995. Motion-Based Recognition: A Survey. *Image and Vision Computing* 13(2): 129-155.
- Chellappa, R.; Wilson, C. L.; and Sirovich, S. 1995. Human and Machine Recognition of Faces: A Survey. *Proceedings of IEEE* 83(5): 705-740.
- Colmenarez, A. J., and Huang, T. S. 1997. Face Detection with Information-Based Maximum Discrimination. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Conference 1997*, 782-787. Washington, D.C.: IEEE Computer Society.
- Darrell, T.; Essa, I.; and Pentland, A. 1996. Task-Specific Gesture Analysis in Real Time Using Interpolated Views. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(12): 1236-1242.
- Darrell, T.; Gordon, G.; Harville, M.; and Woodfill, J. 1998. Multi-Modal Person Detection and Identification for Interactive Systems. In *Proceedings of Computer Vision and Pattern Recognition Conference*. Washington, D.C.: IEEE Computer Society.
- Efron, D. 1941. *Gesture and Environment*. New York: King's Crown.
- Ekman, P., and Friesen, W. 1969. The Repertoire of Nonverbal Behavioral Categories—Origins, Usage, and Coding. *Semiotica* 1:49-98.
- Ekman, P., and Friesen, W. V. 1978. *Facial Action Coding System*. Palo Alto, Calif.: Consulting Psychologists Press.
- Ekman, P.; T. Huang, T.; Sejnowski, T.; and Hager, J., eds. 1993. Final Report to NSF of the Planning Workshop on Facial Expression Understanding. Technical report, National Science Foundation, Human Interaction Lab, University of California at San Francisco.
- Ellis, H. D.; Jeeves, M. A.; Newcombe, E.; and Young, A., eds. 1986. *Aspects of Face Processing*. Zoetermeer, The Netherlands: Martinus Nijhoff.
- Essa, I., and Pentland, A. 1995. Facial Expression Recognition Using a Dynamic Model and Motion Energy. In *Proceedings of the International Conference on Computer Vision*, 360-367. Washington, D.C.: IEEE Computer Society.
- Essa, I., and Pentland, A. 1994. A Vision System for Observing and Extracting Facial Action Parameters. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 76-83. Washington, D.C.: IEEE Computer Society.
- Essa, I.; Darrell, T.; and Pentland, A. 1994. Tracking Facial Motion. In *Proceedings of the Workshop on Motion and Nonrigid and Articulated Objects*, 36-42. Washington, D.C.: IEEE Computer Society.
- Essa, I.; Basu, S.; Darrell, T.; and Pentland, A. 1996. Modeling, Tracking, and Interactive Animation of Faces and Heads Using Input from Video. In *Proceedings of Computer Animation Conference 1996*, 68-79. Washington, D.C.: IEEE Computer Society.
- Etemad, K., and Chellappa, R. 1994. Face Recognition Using Discriminant Eigenvectors. Paper presented at the International Conference on Acoustics, Speech, and Signal Processing, 19-22 April, Adelaide, Australia.
- Face Recognition. 1999. Face Recognition Web Page. Available at www.cs.rug.nl/~peterkf/FACE/face.html.
- Gavril, D. M., and Davis, L. S. 1996. 3-D Model-Based Tracking of Humans in Action: A Multi-View Approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 73-80. Washington, D.C.: IEEE Computer Society.
- Hodgins, J. 1998. Animating Human Motion. *Scientific American* 276(3).
- Kakadiaris, I.; Metaxas, D.; and Bajcsy, R. 1994. Active Part-Decomposition, Shape, and Motion Estimation of Articulated Objects: A Physics-Based Approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 980-984. Washington, D.C.: IEEE Computer Society.
- Kanade, T. 1977. *Computer Recognition of Human Faces*. Cambridge, Mass.: Birkhauser Verlag.
- Kendon, A. 1974. Movement Coordination in Social Interaction: Some Examples Described. In *Nonverbal Communication*. New York: Oxford University Press.
- Kirby, M., and Sirovich, L. 1990. Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. *Pattern Analysis and Machine Intelligence* 12(1): 103-108.
- Krauss, R.; Morrel-Samules, P.; and Colasante, C. 1991. Do Conversational Hand Gestures Communicate? *Journal of Personality and Social Psychology* 61(5): 743-754.
- Kruger, N.; Potzsch, M.; and von der Malsburg, C. 1997. Determination of Face Position and Pose with a Learned Representation Based on Labeled Graphs. *Image and Vision Computing* 15(8): 665-673.
- Kruizinga, P. 1999. Face-Recognition Web Page. Available at www.cs.rug.nl/peterkf/FACE/face.html.
- Landes, B.; Vorbruggen, C.C.; Buhmann, J.; Lange, J.; von der Malsburg, C.; Wurtz, R. P.; and Konen, W. 1993. Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Transactions on Computers* 42(3): 300-311.
- Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago: Chicago University Press.
- Leung, T.; Buri, M.; and Perona, P. 1995. Finding Faces in Cluttered Scenes Using Labelled Random Graph Matching. In *Proceedings of the International Conference on Computer Vision*, 637-644. Washington, D.C.: IEEE Computer Society.

- McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.
- Manjunath, B. S.; Chellappa, R.; and von der Malsburg, C. 1992. A Feature-Based Approach to Face Recognition. In *Proceedings of Computer Vision and Pattern Recognition*, 373-378. Washington, D.C.: IEEE Computer Society.
- Mase, P. 1993a. *ALIVE: An Artificial Life Interactive Video Environment*. In *ACM SIGGRAPH Visual Proceedings*, 189. New York: Association of Computing Machinery.
- Mase, K. 1993b. Recognition of Facial Expressions for Optical Flow. *IEICE Transactions (Special Issue on Computer Vision and Its Applications)* E74(10).
- Moghaddam, B., and Pentland, A. 1995. Probabilistic Visual Learning for Object Detection. In *Proceedings of the International Conference on Computer Vision*. Washington, D.C.: IEEE Computer Society.
- Moore, D.; Essa, I.; and Hayes, M. 1999. Context Management for Human Activity Recognition. Paper presented at the Audio-Visual-Based Person Authentication Conference, 22-23 March, Washington, D.C.
- MPEG. 1999. Moving Picture Experts Group Web Page. Available at www.mpeg.org/.
- Nalwa, V. 1993. *A Guided Tour of Computer Vision*. Reading, Mass.: Addison Wesley.
- Oliver, N.; Pentland, A. P.; and Berard, F. 1997. *LAFER: Lips and Face Real-Time Tracker*. In *Computer Vision and Pattern Recognition*, 123-129. Washington, D.C.: IEEE Computer Society.
- Pelachaud, C.; Badler, N.; and Viaud, M. 1994. Final Report to NSF of the Standards for Facial Animation Workshop. Technical Report, National Science Foundation, Washington, D.C.
- Pentland, A. 1996. Smart Rooms. *Scientific American* 274(4): 68-76.
- Pentland, A.; Moghaddam, B.; and Starner, T. 1994. View-Based and Modular Eigen-spaces for Face Recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 84-91. Washington, D.C.: IEEE Computer Society.
- Phillips, P. J.; Moon, H. J.; Rauss, P.; and Rizvi, S. A. 1997. The FERET Evaluation Methodology for Face-Recognition Algorithms. In *Proceedings of Computer Vision and Pattern Recognition*, 137-143. Washington, D.C.: IEEE Computer Society.
- Phillips, P. J.; Wechsler, H.; Huang, J.; and Rauss, P. J. 1998. The FERET Database and Evaluation Procedure for Face-Recognition Algorithms. *Image and Vision Computing* 16(5): 295-306.
- Picard, R. 1998. *Affective Computing*. Cambridge, Mass.: MIT Press.
- Polana, R., and Nelson, R. 1993. Detecting Activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2-7. Washington, D.C.: IEEE Computer Society.
- Rehg, J. M., and Kanade, T. 1994. Visual Tracking of High DOF Articulated Structures: An Application to Human Hand Tracking. Paper presented at the Third European Conference on Computer Vision, 2-6 May, Stockholm, Sweden.
- Rowley, H. A.; Baluja, S.; and Kanade, T. 1998a. Neural Network-Based Face Detection. *Pattern Analysis and Machine Intelligence* 20(1): 23-38.
- Rowley, H. A.; Baluja, S.; and Kanade, T. 1998b. Rotation Invariant Neural Network-Based Face Detection. In *Proceedings of Computer Vision and Pattern Recognition*. Washington, D.C.: IEEE Computer Society.
- Samal, A., and Iyengar, P. A. 1992. Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey. *Pattern Recognition* 25(1): 65-77.
- Seltz, S., and Dyer, C. 1997. View-Invariant Analysis of Cyclic Motion. *International Journal of Computer Vision* 25(3).
- Sirovich, L., and Kirby, M. 1987. Low-Dimensional Procedure for the Characterization of Human Faces. *Journal of the Optical Society of America* 4(3): 519-524.
- Starner, T. 1995. The MIT Wearable Computing Web Page. Available at wearables.www.media.mit.edu/projects/wearables/.
- Starner, T.; Weaver, J.; and Pentland, A. 1998. Real-Time American Sign Language Recognition Using Desk and Wearable Computer-Based Videos. *IEEE Transactions on Pattern Analysis and Machine Vision* 20(12): 1371-1375.
- Stillman, S.; Tanawongsuwan, R.; and Essa, I. 1999. A System for Tracking and Recognizing Multiple People with Multiple Cameras. Paper presented at the Audio-Visual-Based Person Authentication Conference, 22-23 March, Washington, D.C.
- Swets, D. L., and Weng, J. J. 1996. Using Discriminant Eigenfeatures for Image Retrieval. *Pattern Analysis and Machine Intelligence* 18(8): 831-836.
- Turk, M., and Pentland, A. 1991. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* 3(1): 71-86.
- Visiolonics. 1997. FACET Developer Kit Version 2.0. Visiolonics Corporation. Available at www.visiolonics.com.
- Wiskott, L.; Fellous, J. M.; Kruger, N.; and von der Malsburg, C. 1997. Face Recognition by Elastic Bunch Graph Matching. *Pattern Analysis and Machine Intelligence* 19(7): 775-779.
- Wren, C.; Azarbayejani, A.; Darrell, T.; and Pentland, A. 1997. *PFINDER: Real-Time Tracking of the Human Body*. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7): 780-785.
- Yacoub, Y., and Davis, L. 1994. Computing Spatio-Temporal Representations of Human Faces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 70-75. Washington, D.C.: IEEE Computer Society.
- Yamato, J.; Ohya, J.; and Ishii, K. 1994. Recognizing Human Action in Time-Sequential Images Using a Hidden Markov Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 379-385. Washington, D.C.: IEEE Computer Society.
- Wearable Computing 1999. Wearable Computing Web Page. Available at wearables.www.media.mit.edu/projects/wearables.



Irfan A. Essa is an assistant professor and Imlay fellow in the College of Computing and adjunct assistant professor in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. At Georgia Tech, he is affiliated with the Future Computing Environments effort; the Graphics, Visualization, and Usability Center; and the Intelligent Systems Group in the College of Computing. He has founded the Computational Perception Laboratory that aims to explore and develop the next generation of intelligent machines, interfaces, and environments that can perceive, recognize, anticipate, and interact with humans.

Prior to joining Georgia Tech, Essa was a research scientist in the Perceptual Computing Section of the Media Lab at the Massachusetts Institute of Technology (MIT). He received his Ph.D. from MIT in September 1994. His dissertation dealt with visual analysis and interpretation of facial expressions. He earned his M.S. from MIT (Media Lab and IESL) in 1990 and his B.S. from the Illinois Institute of Technology. His web page is located at www.cc.gatech.edu/~irfan, and his e-mail address is irfan@computer.org.

AL

Direct Control of the Computer through Electrodes Placed Around the Eyes

James Gips^a, Peter Olivieri^a, and Joseph Tecce^b

^aComputer Science Department, Boston College, Chestnut Hill, Mass. 02167

^bPsychology Department, Boston College, Chestnut Hill, Mass. 02167

Abstract

A system has been developed that allows an individual to communicate with a Macintosh computer solely through electrodes placed around his eyes. The user controls the cursor on the screen simply by moving his eyes and head. The user can spell out messages on the Macintosh, play tic-tac-toe, and even play commercial video games by controlling the computer through the electrodes.

1. INTRODUCTION

We are interested in exploring new ways of controlling and interacting with the computer. In this paper we report on experiments to control a computer by utilizing electrical signals recorded through electrodes that are placed on the user's head next to the user's eyes. Electrical signals generated by eye movements then are converted to corresponding movements of a cursor located on the computer screen.

2. BACKGROUND

2.1. Computer Background

The Macintosh II on which this program is run is well over 1,000 times more powerful than the original commercial Univac computer that was offered in 1951, and costs less than 1/1,000 as much in real dollars. The improvements and price cuts in computing power have been continuous over the last 40 years.

What do we do with all of this increasing computer power, all of this capability? Much of the additional capability has been exploited to improve the human-computer interface, to make the computer easier to use, more responsive, and more tightly coupled to the user. The increasing power has allowed for continuing improvement in interface design. In the 1950's, interaction with the computer was in the form of batch processing -- entering information on punch cards and paper tape. In the 1960's, time-sharing and interactive computing with character-oriented

terminals was personal computing. Some graphic graphical user interfaces became popular technologies.

The research in human-computer interaction for direct input

2.2. Electro

There are electrodes, it is muscles (EMG).

When placed on the eye, electrodes measure the potential between the eye and the electrodes placed on the eye relative to the electrode placed on the leg. For every degree of signal, [1] Electrodes are used in clinical work, such

3. RELATED

There are two ways to monitor eye movements: electrodes on the

3.1. Eye tracking

The most common method is to monitor the subject's eye movements by a camera. The subject is monitored by a camera and finds the computer then camera. Potter, advertising, target disabled. [4,5]

3.2. Using Electrodes

A recent trend has been to use electrodes on the face to monitor eye movements.

terminals was developed. In the 1970's, we saw the introduction of the first personal computers, which were predominantly character-oriented but included some graphics and inputs through paddles and joysticks. In the 1980's, the graphical user interface using icons, graphics and a mouse or trackball for input became popular. In the early 1990's we are seeing a proliferation of new interface technologies: voice input, handwriting input, datagloves, goggles, and multimedia.

The research described in this paper seeks to explore ways to expand the human-computer interface. In particular, we are interested in the use of electrodes for direct input of information into the computer.

2.2. Electrophysiology Background

There are many sources of electrical signals in the human body. Through electrodes, it is possible to measure signals from the heart (EKG), from skeletal muscles (EMG), and from the brain (EEG).

When placed around the eyes in the work reported here, the electrodes measure the electro-oculographic potential (EOG), the variation in the standing potential between the retina and the cornea, which is a function of the position of the eye relative to the head. The potential difference measured between electrodes placed above and below the eye indicates the vertical position of the eye relative to the rest of the head. The potential difference between electrodes placed to the left and right of the eyes indicates the horizontal position of the eyes. For every degree of angle there is a change of approximately 20 microvolts in the signal. [1] Electrophysiology recordings around the eye have several uses. They are used to monitor eye movements, including eyeblinks [2], and are used in clinical work, such as in the diagnosis of vestibular disease [3].

3. RELATED WORK

There are two areas of related work. The first area concerns other approaches to monitor eye movements. The second area concerns other approaches that use electrodes to control the computer directly.

3.1. Eye tracking through cameras

The most common method of tracking eye movements is by camera. The eye of the subject is illuminated by a low level infrared light while the eye area is monitored by an infrared sensitive video camera. A computer analyzes the image and finds the pupil of the eye, which appears dark in the illumination. The computer then calculates the pupil size and position coordinates relative to the camera. Potential application areas include monitoring customer responses to advertising, target tracking, virtual reality, and facilitating computer control by the disabled. [4,5]

3.2. Using EEG to control the computer

A recent front page article in the New York Times [6] reported on various efforts over the past decade to develop systems that can be controlled by users through electrodes monitoring EEG ("brain waves"). Perhaps this could lead to the ideal

system: DWIW -- Do What I Wish. No need to type or use the mouse or even move your eyes. Your very wish is the computer's command. Unfortunately, this is not very easy to accomplish.

Farwell and Donchin [7] developed a system for communicating with a computer through EEG. The subject stares at a letter in a grid. The computer repeatedly flashes columns and rows of the grid and determines the cell in which the subject is staring by analyzing the resulting EEG signal.

Wolpaw et. al. [8] developed a system that allows the user to move a cursor in one direction up or down a screen by learning to change the mu rhythm amplitude. The mu rhythm is the component of the EEG signal between 8 and 12 Hz.

Sutter has developed the most advanced system for controlling the computer through EEG. Sutter's system displays a grid of letters on the screen, each in turn. The system "is based on the fact that objects in the center of the field of vision generate a larger EEG response than those at the periphery." [9] Thus, the system monitors the EEG after each letter is displayed and determines which letter generated the largest EEG. This is the letter the user was looking at. Sutter developed the system as a way to help severely disabled people communicate.

4. METHODS

With our system, the user sits down in front of the computer. Five electrodes are placed on his head. An electrode is placed 1 cm. above the right eyebrow and another 2 cm. below the right eye. Electrodes are placed 2 cm. to the right and left of the outer canthi. A clip electrode is attached to the user's ear to serve as a ground. (See Figure 1)



Figure 1. Placement of electrodes. One more electrode is next to the left eye.

The leads from which amplify the signal to an analog-to-digital converter board in a Macintosh for data acquisition.

Two different testbed environments are displayed as if a calibration scheme is used to control the cursor.



Figure 2. Control.

In particular, the user interacts with the system (Figure 3) with no physical electrodes. In the testbed, the black square cursor is moved so that if the cursor is selected and is added to the user's spelling.

Similarly, another testbed environment is used to control the computer by moving the cursor to place an X or O. This requires sufficient time.

mouse or even move
fortunately, this is not

communicating with a
grid. The computer
lines the cell in which

er to move a cursor in
mu rhythm amplitude.
3 and 12 Hz.

controlling the computer
a screen, each in turn.
of the field of vision
[9] Thus, the system
etermines which letter
as looking at. Sutter
ople communicate.

ir. Five electrodes are
he right eyebrow and
cm. to the right and left
er's ear to serve as a

The leads from these electrodes are connected to two Grass 7P122B amplifiers, which amplify the signals by a factor of 5,000. The amplifier outputs are connected to an analog-to-digital board in a Macintosh computer. Currently we are using two different Macintosh systems: A GW Instruments A/D converter and data acquisition board in a Macintosh IIx computer, and a National Instruments A/D converter and data acquisition board in a Quadra 950 Macintosh computer system.

Two different software systems have been developed. The first is a custom testbed environment. In one mode of this software, the incoming signals are displayed as if on an oscilloscope. We can try various signal conditioning and calibration schemes. One part of the program allows us to use the incoming signals to control the cursor on the screen to spell out messages. The user can move the cursor around the screen by moving his eyes and head (Figure 2).



Figure 2. Controlling the cursor on the screen.

In particular, the user can spell out a message by looking at letters in a grid (Figure 3) with no other input to the computer except what is coming through the electrodes. In the screen in Figure 3, the user is controlling the position of the black square cursor by moving his eyes and head. The machine is programmed so that if the cursor stays in a box for at least half of a second, then that letter is selected and is added to a message at the bottom of the screen. In the illustration, the user has spelled out "HELLO EVERYON" and is about to add the "E".

Similarly, another part of the program allows the user to play tic-tac-toe with the computer by moving his eyes and head to look at the square where he wants to place an X or O. The square will be selected if the cursor remains in the square for sufficient time.

ext to the left eye.

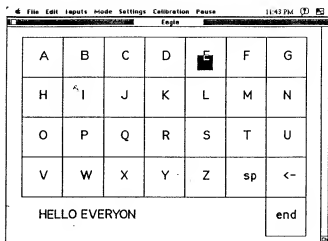


Figure 3. Spelling out a message on the screen.

The second program is a utility developed by Aaron Walsh of Mantis Software that substitutes the values received from the analog to digital converter directly into the mouse coordinates in system memory. Thus the user can run standard Macintosh software by moving his eyes and head. This adds a new dimension to video games. For example, instead of using the mouse to drive the car in the video game, the user can drive it simply by looking to the left or right.

5. RESULTS

Half a dozen people have used the system during its development. A new user takes a few minutes to get used to the system and learns that the signal can be changed by moving the eyes while keeping the head relatively still or by moving the head while keeping the eyes relatively fixated or, more naturally, by a combination of eye movements and head movements. The signal indicates the relative position of the eyes in the head. The system basically tracks eye movements. However, nonlinearities or drifts in the signal can be accommodated by slight movements of the head. After a session or two the whole process becomes automatic (and fun!) and performance in spelling out messages is close to flawless. (A "delete key" in the grid even allows for error correction, if necessary.) In a preliminary attempt to measure best performance, a 20 character message was spelled out in 21 seconds. Part of the time in selecting a letter is required to sweep the cursor to the desired grid cell, part to hold the cursor in the cell long enough to specify the letter.

6. FUTURE WORK

Work is progressing on miniaturizing the electrophysiology exploring options for working on algorithms to better placement on the technology. vehicles both sir

7. ACKNOWLEDGMENTS

We would like to thank the software module College undergrad

8. REFERENCES

1. B. Tursky, J. Patterson (eds), 1974.
2. J. J. Tecce, Yearbook of "Clinician", T
3. C. W. Stock, September
4. R. Razdan, September
5. G. A. Myers, February 9,
6. A. Pollack, February 9,
7. L. A. Farwell, and Clinical
8. J. Wolpaw, computer in Neurophysic
9. R. B. Duffy, "

6. FUTURE WORK

Work is proceeding on three fronts. We are exploring various options for miniaturizing the hardware required. Right now we are using rack mounted electrophysiology amplifiers, so we are tethered to stationary equipment. We are exploring options for portable, preferably wearable, amplifiers. We also are working on algorithms for dynamic calibration and tracking in the software to allow us to better and more accurately translate the electrical signals into cursor placement on the screen. Of course, we also are exploring application areas for this technology. Virtual reality, exploration of ultralarge databases, control of vehicles both simulated and real, video games, all seem likely candidates.

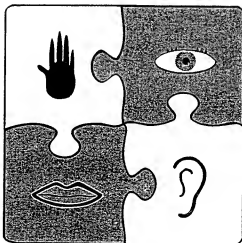
7. ACKNOWLEDGMENTS

We would like to thank Aaron Walsh of Mantis Software for developing the software module and Jason McHugh and Donald Green, two very talented Boston College undergraduate students, for their assistance on this project.

8. REFERENCES

1. B. Tursky, "Recording Human Eye Movements", in R. Thompson and M. Patterson (eds.), *Bioelectric Recording Techniques*, Part C, Academic Press, 1974.
2. J. J. Tecce, "Psychology, physiological and experimental", McGraw-Hill Yearbook of Science and Technology, 1992, pp. 375-377.
3. C. W. Stockwell, "Computerized Vestibular-Function Tests: An Overview for the Clinician", *The Hearing Journal*, November 1988, Vol. 41, No. 11.
4. R. Razdan and A. Kielar, "Eye Tracking for Man/Machine Interfaces", *Sensors*, September 1988.
5. G. A. Myers, "The EyeMouse", *T.H.E. Journal*, January 1992, pp. 13 - 15.
6. A. Pollack, "Computers Taking Wish as Their Command", *New York Times*, February 9, 1993, p. 1.
7. L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials", *Electroencephalography and Clinical Neurophysiology*, 1988, Vol. 70, pp. 510-523.
8. J. Wolpaw, D. McFarland, G. Neat, and C. Forneris, "An EEG-based brain-computer interface for cursor control", *Electroencephalography and Clinical Neurophysiology*, 1991, Vol. 78, pp. 252-259.
9. R. B. Duffy, "PC Mind Control", *PC-Computing*, November 1989, pp. 155-156.

A.M



ISAAC
DUBLIN 1998

24-27th August 1998,
UCD, Dublin,
Ireland

Proceedings

isaac
International Society for
Augmentative and Alternative Communication

Progress with EagleEyes

James Gips, Computer Science Department
Philip DiMattia, School of Education
Francis X. Curran, School of Education

Boston College
Chestnut Hill, Mass. 02167 USA
gips@bc.edu
www.cs.bc.edu/~gips

EagleEyes is a technology that allows a person to control the computer through electrodes by moving his or her eyes and head. EagleEyes works as a general mouse-replacement device on a Windows or Macintosh computer. EagleEyes allows people who have no voluntary muscle control below the neck and cannot speak to communicate through art, music, and by clicking out messages from choices on the screen.

How does EagleEyes work?

EagleEyes is based on measuring the EOG or electro-oculographic potential, a small electrical potential which indicates the position of the eye relative to the head. Surface electrodes are placed on the user's head, above and below one eye, and on the temples to the left and right of the eyes. The five electrodes are connected to two electrophysiological amplifiers which are connected to a data acquisition board in the computer. A program translates the signals received from the electrodes into the position of the cursor on the screen. When the user changes the position of the eyes relative to the head, the cursor moves. Basically, the cursor follows the location that the user is looking at on the screen. The eyes replace the mouse. Selection is made by holding the cursor in a small area of the screen for a short period of time, which causes a mouse click.

Learning to use EagleEyes

Using EagleEyes is an acquired skill. Boston College undergraduates learn to use the system well enough within an hour to use EagleEyes to spell out messages at the rate of one character every 2.5 seconds. People with profound disabilities can require much more time to learn to use the system.

Who has used EagleEyes?

Over the past two years we have been teaching some severely disabled children and young adults from the Greater Boston area on the technology. These children are able to "eye paint" (create "finger paintings" by moving their eyes), compose music, and run educational, entertainment, and communications software just by moving their eyes. We have had some exciting successes. We

have established our first satellite facility in Middleboro, Mass. Two students with severe cerebral palsy have advanced to the point where they currently are using EagleEyes for cognitive academic activities in their school programs. Both of these students now have EagleEyes systems in their homes.

At the invitation of the California Pacific Medical Center in San Francisco we tried EagleEyes with ten people with ALS. In a 60 to 90 minute session each learned to use the system well enough to play video games (hit at least 7 out of 10 aliens). Five learned to use EagleEyes well enough to spell out messages. One gentleman with advanced ALS spelled out the message "There is no way to the end of the journey but to travel the road that leads to it"

Communicating through Painting and Eye Writing

One of the first programs we use with children with disabilities is a custom-developed "paint" program that puts thick lines of color on the screen wherever the child moves the cursor by moving the eyes. These children are not able to fingerpaint. We print out these "eye paintings" on a color printer and give them to parents to hang on their refrigerators or walls. Undergraduate assistants have learned to write out their names and other short messages with their eyes using the paint program.

Communicating through Music

The Axe is an interactive music program developed by Harmonix Music Systems, a spin-off of the MIT Media Lab. Through EagleEyes people can use The Axe to compose and play music in real time just by moving their eyes.

Communicating through Language

EagleEyes works with commercial communications programs, such as Speaking Dynamically.

We have developed various spelling programs that allow people to use EagleEyes to spell out messages letter by letter. The appropriate spelling program to use depends on the person's skill in controlling the cursor through EagleEyes. A person with good control can use a full 35 character keyboard on the screen. A person with poorer control might need a two-level system with six possible choices or a yes/no system with two choices at a time.

Communicating with Mobile Devices

We have developed and publicly demonstrated EagleEyes controllers for a toy car and for two powered wheelchairs. The idea is for the person to look in the direction where the mobile device is to move: up to go forward, left to go left, down to go backwards, right to go right, at an area in the middle to stop. One of the wheelchairs, Wheelasley, is an experimental robotic system with its own sensors and onboard computer that allow it to travel semi-autonomously and automatically avoid hitting obstacles.

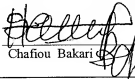
IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Gips and Betke)
Serial No: 09/892,254) Group Art Unit: 2173
Filed: June 27, 2001) Examiner: To Be Assigned
Title: Automated Visual Tracking For Computer Access)

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as First Class Mail, postage prepaid, in an envelope addressed to: Commissioner for Patents, Washington, D.C. 20231.

03/29/02
Date of Signature
and of Mail Deposit


Chafiu Bakari

Commissioner for Patents
Washington, D.C. 20231

**SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT BASED ON
INTERNATIONAL SEARCH REPORT IN A RELATED PCT APPLICATION**

Sir:

In compliance with the requirements of 37 C.F.R. 1.56, submitted herewith on Form PTO-1449 is a list of publications identified in an International Search Report in a related PCT application; a copy of each publication is being submitted herewith. Applicants respectfully request that the Examiner consider the listed documents and indicate that they were considered by making appropriate notations on the attached Form 1449.

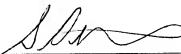
This submission does not represent that a search has been made or that no better art exists. Nor does it constitute an admission that the listed documents are material or constitute "prior art." If the Examiner applies the cited documents as prior art against any claim in the

application and Applicants determine that the cited documents do not constitute "prior art" under United States law, Applicants reserve the right to present to the Office the relevant facts and law regarding the appropriate status of said document. Applicants further reserve the right to take appropriate action to establish the patentability of the disclosed invention over the listed documents, should one or more of the cited references be applied against the claims of the present application.

This Information Disclosure Statement is being filed within three months of the mailing date of the International Search Report; therefore, no fee is believed to be due in connection with the filing of this disclosure. However, the Commissioner is authorized to credit any overpayment or charge any deficiencies to/from our Deposit Account, No. 06-1448.

Respectfully submitted,

FOLEY, HOAG & ELIOT, LLP

By: 

Stephen B. Deutsch
Attorney for Applicants
Reg. No. 46,663

Dated: 03/29/02
Customer No.: 25181
Patent Group
Foley, Hoag & Eliot LLP
One Post Office Square
Boston, MA 02109

Voice: (617) 832-1000
Facsimile: (617) 832-7000

**INFORMATION DISCLOSURE CITATION
IN AN APPLICATION**
(Use several sheets if necessary)

Docket Number (Optional)
BOK-002.01 (00888-201)

Application Number
09/892,254

Applicant
Gips and Betke

Filing Date
June 27, 2001

Group Art Unit
2173

U.S. PATENT DOCUMENTS

EXAMINER INITIAL	DOCUMENT NUMBER	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
AN	US 4,975,960	12/04/90	Petajan	381	43	

FOREIGN PATENT DOCUMENTS

	DOCUMENT NUMBER	DATE	COUNTRY	CLASS	SUBCLASS	Translation	
						YES	NO

OTHER DOCUMENTS
(Including Author, Title, Date, Pertinent Pages Etc.)

AO	Arabnia R. H., "A Computer Input Device for Medically Impaired Users of Computers" Proceedings of the Johns Hopkins National Search for Laurel, MD, USA, 1-5 Feb. 1992, Los Alamitos, CA, USA, IEEE Comput. Society, US, PP. 131-134, (1992)
AP	Ohtani et al.; "A Pointing Device Using Coordinate Transformation of Neurofuzzy GMDH", 1998 Second International Conference on Knowledge-Based Intelligence Electronic Systems Proceedings Kes, Adelaide, Australia, IEEE New York, pp. 108-115, (21-23 April 1998)
AQ	Sako et al.; "Real-Time Facial- Feature Tracking Based on Matching Techniques and Its Applications", Pattern Recognition, Vol. 2 : 320-324, (October 9, 1994)
AR	Takami et al.; "Computer Interface to Use Head and Eyeball Movement for Handicapped People", IEEE International Conference, Vancouver, CANADA, BC, pp. 1119 -1123, (October 22, 1995)
AS	International Search Report Completed on January 16, 2002 and Mailed on February 04, 2002.
EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP § 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to the applicant.

A Computer Input Device for Medically Impaired Users of Computers

Hamid R. Arabnia

The University of Georgia, Department of Computer Science

Abstract

In this paper, we introduce a novel computer input device capable of receiving input by processing the images of the scene viewed by a camera; i.e., there is no need for any physical contact. The user of this input device can use any part of his body to issue a computer command. The system described here is a cursor-positioning input device. In addition to cursor-positioning commands, it also provides a simulated keyboard to allow users input text without using the physical keyboard. The system described here has been successfully built; our preliminary experience in using the system has been very encouraging.

1: Introduction

There are two ways in which an interactive user of a computer can give commands to his computer:

- (a) Via an input device which needs to be physically touched by the user.
- (b) Via an input device which receives input with no physical contact.

Devices in class (a) include: the keyboard, joystick, and mouse. These devices are widely used and have always been the principal sources of input to a computer system. The implementations (software and hardware) of these systems are well understood.

There are only two examples in class (b): the voice recognition and the eye-tracking systems. The input devices in this class are quite natural for us to use but the systems built so far are quite limited. The eye-tracking system is a cursor-positioning input device. It allows users to select computer commands by

looking at the command in a display menu. Unfortunately, to a large extent this device is still an experimental system. The voice recognition device has been used more widely. It processes sound waves in order to recognize the input (command). Thus, the user issues the command by voicing it. While the user is reading the command the device processes the sound waves generated by the user in order to recognize the command.

The computer input device presented in this paper is novel; it has been successfully built and is operational. The device is used to select a command that is displayed in a menu; the user interface incorporates a window popup and menu pulldown system. Therefore, the system is a cursor-positioning input device (like a mouse or an eye-tracking system). The device can be regarded to be a system in class (b). It receives input through a camera; therefore, no contact with any physical device is required in order to issue a command. It determines what command is issued by processing the images viewed by the camera. The system achieves this in real time by performing only simple low-level image processing operations. All the hardware components of the device have been purchased off the shelf and so the cost of building the system has been very reasonable. The device introduced here is referred to as *vision-based input device*.

The user of the vision-based input device can use any part of his body to issue a computer command. This makes the system potentially a valuable tool for medically impaired users of computers. The importance of having computer input devices designed for disabled users of computers has been realized recently: the "Rehabilitation Act" which was amended by the U.S. Congress to cater for disabled users of computers, attests to this (refer to [1] for an interpretation of the act).

The vision-based computer input device introduced here, has a great potential to be useful to physically disabled users of computers. The system can be used by people with *different degrees of physical disability*. This feature of the device is very significant since this will allow the system to be mass produced in lower the cost. Compared to the eye-tracking and voice recognition systems, the input device introduced here is very simple in both hardware and software. It is this simplicity which makes the system attractive to us.

2: Hardware

The major components of the vision-based input device are: a monitor, a simple image capturing board [2], a CCD camera, and the supporting software and programs which are developed by us. These components are interfaced to a standard IBM compatible Personal Computer (a 286 system with 640 KBytes main memory) running under the DOS operating system. The PC used is self-contained (i.e., it has its own monitor, hard disk, ...). Therefore, there are two monitors in our current set up. One is referred to as the *camera monitor* which is used to display the images viewed by the camera and the other is referred to as the *PC monitor* which is used to display text and application menus. The set up is shown in Fig. 1. All the hardware components have been purchased off the shelf. Unlike other input devices designed for disabled users of computers, the vision-based device can be mass produced to lower the cost. It can be mass produced because the system is not targeted at any particular disability; it has the potential to be useful to people with different degrees of disability.

3: The vision-based device

When the vision-based device is in use, five boxes are displayed on the *camera monitor* together with the image sensed by the camera. These boxes are displayed in a transparent form. Each box represents a cursor movement command (move up, move down, move right, move left, and select command). The function of the command that a box represents is identified by a symbol displayed inside the box: ↑ for move up, ↓ for move down, → for move right, ← for move left, and SL for selecting a command (refer to Fig. 1).

The user can interactively select one of those five commands by moving and positioning a part of his body so that the image of that part overlaps the box he wishes to select (Fig. 1 shows an example in which a foot is used; it can be ANY part of the body). When a box is selected, the corresponding cursor movement command is executed on the *PC monitor*. The boundary of the selected box will change color highlighting the command box (see Fig. 1). For example, in order to move a cursor that is displayed on the *PC monitor* up, the user has to position the image of a part of his body so that the image of that part overlaps the box that represents the "move up" command (i.e., first box from the left that is displayed on the *camera monitor* — see Fig. 1). The cursor will stop moving when the image of the part of the body that is used to select the command is removed from the command box. In this way, the user can select any command within an application menu displayed on the *PC* or the *camera monitor*.

Elsewhere [3] we have reported how DOS commands are issued using our vision-based input device and showed the DOS command menus of our system. The DOS commands are under the command box *cmd* as submenus (refer to Fig. 1). Details on how DOS commands are selected using the vision-based input device can be found in [3].

4: A simulated keyboard

The system also provides a simple simulated keyboard which enables the user to input characters without using the computer keyboard. The keyboard is displayed on the frame buffer of the image capturing board; thus it appears on the *camera monitor*. The simulated keyboard has a tree structure. Since the layout of the simulated keyboard is multi-level, the complete character set cannot be viewed in one place. However, the user can easily follow the hierarchy of the tree by choosing the desired tree path which contains the character (letter, digit, ...) he wishes to select. The structure of the simulated keyboard is shown partially in Fig. 2 through Fig. 5. Fig. 2 shows the top level of the simulated keyboard; this will be displayed on the *camera monitor* when it is activated. At this level, if the user selects the command *Exit*, the simulated keyboard will be exited (i.e., the main menu of the input device will be activated). The command *Erase* deletes the last character keyed in. *Space* is used to key in the space character.

Return is the new line character. *Other* contains all the punctuation symbols as its children. *Digit* contains the numeral symbols (0 to 9) as its children. Finally, *Letter* contains all the alphabet set as its children (A to Z and a to z). Fig. 3 shows the children of *Letter*; this menu will be displayed on the *camera monitor* when *Letter* is selected. At this level, if *Exit* is selected, the system will display the previous level of the tree (i.e., it displays Fig. 2). The children of box 'A — E' are the characters A to E as shown in Fig. 4. Similarly, the children of box 'E — I' are the characters E to I as shown in Fig. 5. The remaining children of *Letter* are the rest of the alphabet character set (not shown in the figures).

The same technique is used to provide other symbols (digits, punctuations, ...).

5: Using the simulated keyboard

As an example, suppose the user of the system wishes to use the simulated keyboard to key in characters. When the vision-based device is first activated, the main menu (shown in Fig. 1) will be displayed on the *PC monitor*. The user then selects the command *kbd* (the box on the far right of the *PC monitor*). This activates the simulated keyboard. The menu shown in Fig. 2 will be displayed on the *camera monitor*. Suppose the user wishes to key in the character 'F'. He would move the image of a part of his body so that the image of that part overlaps the box named *Letter*; i.e.,

he selects *Letter*. The system will display the menu shown in Fig. 3. The user now needs to select the box named 'E — I'; when selected, the system will display the menu shown in Fig. 5. Finally, he selects the box named *F*; when selected, the character will be displayed on the *PC monitor* to show the selection. In this way the user can key in any characters. This may appear to be too troublesome, but it is quite easy to use in practice. With only three strokes the user can select any letter he wishes to key in.

It is important to note that although the vision-based input device provides a simulated keyboard, it is intrinsically a cursor-positioning input device.

References

- [1] Ladner R. E., "Computer Accessibility For Federal Workers With Disabilities: Is The Law", *Communications of The ACM*, vol. 32, no. 8, pp. 952-956, 1989.
- [2] Data Translation Inc., "Preliminary User Manual for DT2803 — DT2803 Video I/O subsystem for IBM Personal Computers", 1984.
- [3] Arabnia H. R. and Chen C. Y., "A Remote Media Vision-based Computer Input Device", *The Proceedings of The Visual Communications and Image Processing 91: Image Processing — The International Society for Optical Engineering (SPIE)*, Boston, Nov., Edited by Kou-Hu Tzou and Toshio Koga, vol. 1606, part 2 of 2, pp. 917-926, 1991.

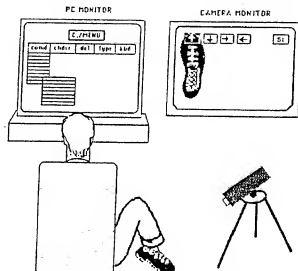


FIG. 1. The vision-based input device

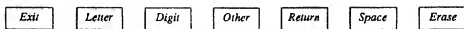


FIG. 2. The top level of the simulated keyboard

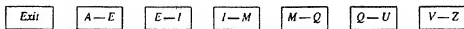


FIG. 3. The children of *Letter* (see Fig. 2)

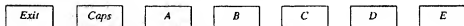


FIG. 4. The children of 'A—E' (see Fig. 3)

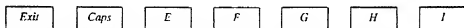


FIG. 5. The children of 'E—I' (see Fig. 3)

ED. 21-04-1998

P. 108-115

8

A Pointing Device Using Coordinate Transformation of Neurofuzzy GMDH

Takashi Ohtani
Hidetomo Ichihashi
Tetsuya Miyoshi
Naoki Tani

College of Engineering
Osaka Prefecture University
1-1, Gakuencho, Sakai, Osaka, 599-8531
JAPAN

Keywords: Pointing Device, Model Selection, RBF, Neurofuzzy, GMDH

Abstract

Personal computers have potential possibility as a tool for aiding users who have disabilities. Though many computer input devices for the use of the handicapped have been developed, they do not work well when they are set away from the user or in aslant. We have developed a head-controlled pointing device that translates handicapped person's movements into direct movements of the computer's cursor by measuring several color markers on the user's head. For sensitivity tuning of the cursor movement and coordinate transformation, the neurofuzzy GMDH (Group Method of Data Handling) is employed, whose building blocks are represented by RBFs. A heuristic model selection criterion called "Distorter" is employed to determine the optimum number of layers in neurofuzzy GMDH. The click motion is replaced with opening the mouth. The software is developed for the use of almost all MS-Windows95 application programs.

1 Introduction

Personal computers have potential possibility as a tool for aiding users who have disabilities. The spread of the internet has brought about opportunities for increasing technical knowledge and presenting art, literature, computer software and other types of materials.

Recently, the pointing devices such as the mouse, trackball and slidepad play an important role under the Graphical User Interface (GUI) of Operating System (OS). For example, most Windows program can be operated by using the keyboard and mouse. Head-controlled interfaces provide an alternative method to accessing a computer and have been beneficial for some individuals with impaired upper-extremity control caused by spinal cord injury degenerative muscular conditions, or cerebral palsy. Many head-controlled interfaces such as HeadMaster (Prentke Romich Company), HeadMouse (Origin Instrument Corp.) and

TRACKER (Madanta Communication Inc.) for use by the handicapped has been developed. These pointing devices translate movement of the users' forehead into direct movement of the computer's cursor by using supersonic wave, optical or laser sensors.

The control unit or sensor of laser must be centered on top of and in the same vertical plane as the CRT. If not, cursor may not move according to the movement of user's head. Furthermore, the marker may disappear if the control unit or sensor is placed away from the CRT or aslant by a bedridden user.

We propose a new type of pointing device which tracks the cursor without using hand, where the neurofuzzy (NF-)GMDH [13-16] is adapted for the coordinate transformation, whose building blocks are represented by RBF networks [10,11,17]. A heuristic model selection criterion [16] called "Distorter" is employed to determine the optimum number of layers in NF-GMDH. The user can place the camera arbitrarily and the cursor movement sensitivity is adjusted as the user likes. The click motion is replaced with opening the mouth. The software is developed for the use of almost all MS-Windows95 application programs.

2 Equipments

The system consists of a CCD camera, Color Multi Tracker (Emtech Inc.), a monitor and a personal computer. The computer is connected to the tracker by a GP-IB (General Purpose Interface Board). The measured data is transferred to the computer and the cursor is displayed on the CRT. Following the movement of the head, the cursor moves, and the

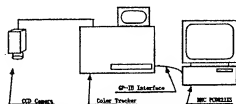


Figure 1: System configuration

click and drag functions are also realized. Fig.1 shows the system configuration.

We develop the system by Microsoft Visual C++ Version 4.0 Standard Edition, BORLAND C++ FOR WINDOWS Version 3.1, Microsoft Visual Basic Version 2.0 Professional Edition. The experimental task was programmed and performed on a NEC PC-9821XA personal computer with a Pentium 90MHz processor.

3 Coordinate Transformation by Neurofuzzy GMDH

We employ the neurofuzzy (NF-)GMDH [13-16] for the coordinate transformation. In [15], it is shown that NF-GMDH with conventional backpropagation learning can converge as fast as the neural network with some accelerated learning method [2].

3.1 Brief Survey of NF-GMDH

NF-GMDH is a kind of Adaptive Learning Network (i.e., a network type of GMDH) [3,12] in the hierarchical structure, whose building blocks are represented by RBF networks. RBF is reinterpreted as both a simplified fuzzy reasoning model and as a three layered neural network [1,5,18].

Fig.2 shows an example of the 4 layered NF-GMDH model in which each layer has 3 building blocks. The inputs of the m -th model in the p -th layer are the output variables y^{pm} of the $(m-1)$ -th and the m -th models in the $(p-1)$ -th layer, then $x_1^m = y^{p-1,m-1}$, $x_2^m = y^{p-1,m}$, where $m-1$ is M when $m=1$.



Figure 2: Structure of Neurofuzzy GMDH with six input variables

Let A_{ki} denote the membership function of the k -th rule ($k = 1, 2, \dots, K$) in the domain of i -th input variable x_i^{pm}

$$A_{ki}(x_i^{pm}) = \exp \left\{ - \frac{(x_i^{pm} - a_{ki}^{pm})^2}{b_{ki}^{pm}} \right\} \quad (1)$$

where, a_{ki}^{pm} and b_{ki}^{pm} are unknown parameters.

The compatibility degree μ_k^{pm} of the premise part of the k -th rule, which infer output y^{pm} of p -th layer and m -th model, is computed with the algebraic product operation as:

$$\mu_k^{pm} = \prod_{i=1}^3 A_{ki}(x_i^{pm}) \quad (2)$$

Let the conclusion part of the fuzzy inference rule be simplified as a real number w_k^{pm} , we have the output p^{pm} of this fuzzy model as:

$$y^{pm} = \sum_{k=1}^K \mu_k^{pm} w_k^{pm} \quad (3)$$

Eq.3 can be regarded as the Gaussian RBF network which can be capable of approximating any continuous mappings within an arbitrary accuracy.

The final output y is given by the average of outputs in the last layer (P -th layer).

$$y = \frac{1}{M} \sum_{m=1}^M y^{pm} \quad (4)$$

Let y^* be the target value and the performance index of the error be

$$E_1 = \frac{1}{2} (y - y^*)^2 \quad (5)$$

The learning is successively proceeded to minimize E_1 by using gradient of E_1 with respect to the unknown parameters. The learning rule, based on the least mean square approach (the gradient descent method) is derived as follows:

$$w_k^{pmNEW} = w_k^{pmOLD} - \tau \mu_k^{pm} \delta^{pm} \quad (6)$$

$$a_{k,i}^{pmNEW} = a_{k,i}^{pmOLD} - \tau \mu_k^{pm} w_k^{pm} \frac{2(x_i^{pm} - a_{k,i}^{pm})}{b_{k,i}^{pm}} \delta^{pm} \quad (7)$$

$$b_{k,i}^{pmNEW} = b_{k,i}^{pmOLD} - \tau \mu_k^{pm} w_k^{pm} \frac{2(x_i^{pm} - a_{k,i}^{pm})^2}{(b_{k,i}^{pm})^2} \delta^{pm} \quad (8)$$

where, τ is a positive small constant and δ^{pm} is the backpropagation error. δ^{pm} is $\delta^{pm} =$

$(y - y^*)/M$ for the last layer and

$$\delta^{pm} = 2 \sum_{l=m}^{m+1} \delta^{p+1,l} \times \sum_{q=1}^4 w_q^{p+1,l} \mu_q^{p+1,l} \frac{(y^{pm} - a_{q,r}^{p+1,l})}{b_{q,r}^{p+1,l}} \quad (9)$$

for intermediate layers, where $r = 1$ if $l = m+1$, and $r = 2$ if $l = m$, and $m+1$ equals 1 if $m = M$. It is assumed that the input/output pairs of data are normalized in unit interval $[0,1]$.

The above operation is performed for each training data, which thus implies one iteration of learning. The operations are repeated for a fix number of learning sessions.

3.2 Model Selection

In conventional GMDH [6], several heuristic selection criteria are developed in order to choose the optimum model to the purpose of modelling from among competing models with different complexity [7]. The model must not be determined only by minimizing sum of square error. For this problem, the heuristic selection criteria are developed in GMDH algorithm, which combined with regulation called "the principle of external complementation". Regularity criteria [8] are based on the heuristic hypothesis, which the model identified for one part of sample data should be as similar to one for another part as possible. However, these criteria are pointed not to be effective by experimental investigation for the structure identification, and are not robust to large level of noise. The unbiasedness criteria [7] (i.e., unbiasedness of solutions and unbiasedness of coefficients) are directed, which are based on the heuristic hypothesis that the model with the true structure is least sensitive to changes of the original information with comparatively small noise.

Cross validation technique [9] divides sample data into two parts training and checking, and estimates the parameters of a model. However, a division of the sample data decreases the reliability of parameters when the number of sample data is comparatively smaller than the number of parameters within a model. In order to eliminate this effect, minimum bias cri-

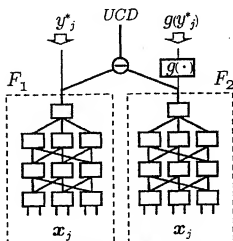


Figure 3: Conceptual figure of the model selection by the Distorter

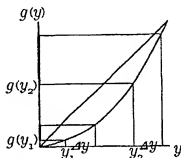


Figure 4: Nonlinear transformation by the Distorter $g(\cdot)$

teria were developed which do not require the division of data. Among others the unbiasedness criterion with Distorter (UCD) [16] is a preferred one which estimates parameters of a model by using all of given data and selects an optimum model.

Fig.3 shows the structure of NF-GMDH with the distorter. In the figure, there are two hierarchical models with the same structure. To identify the model F_1 , following loss function E_1 on the j -th input/output pair is

successively minimized.

$$E_1 = \frac{1}{2} \{y_j^* - F_1(x_j)\}^2 \quad (10)$$

The model F_2 is equipped with the distorter which is a nonlinear transformation function $g(\cdot)$. To identify the model F_2 , following loss function E_2 on the j -th input/output pair is successively minimized.

$$E_2 = \frac{1}{2} \{g(y_j^*) - g(F_2(x_j))\}^2 \quad (11)$$

$g(\cdot)$ is chosen from monotonic functions whose input and output take their values in interval $[0,1]$. Non-monotonic functions are not suited for the learning because it may happen that $g(F_2(x_i)) = g(F_2(x_j))$ in spite of $F_2(x_i) \neq F_2(x_j)$.

The learning rule of the model F_2 is almost same as that of the model F_1 . Only the back-propagation error δ^{pm} in the last layer distinguishes them, that is,

$$\delta^{pm} = -\frac{1}{M} \{g(y_j^*) - g(F_2(x_j))\} \times g'(F_2(x_j)) \quad (12)$$

Eq.12 suggests that the linear function is not suited for $g(\cdot)$ because δ^{pm} of model F_2 becomes the constant multiple of that of the model F_1 and employing any linear transformations are equivalent to changing the learning constant among model F_1 and F_2 . On the other hand, if we assume $g(y) = y^2$ and $y_2 > y_1$, then $g(y_2 + \Delta y) - g(y_2) > g(y_2 + \Delta y) - g(y_1)$. Fig.4 shows this situation in which two model output y_1 and y_2 are assumed to have same error Δy but the backpropagation errors are different. In model F_2 , parameter updates depend on the magnitudes of model output if the error Δy between the target value and the model output are the same. In this case the errors of the data whose target values are close to 1 decrease faster than the error of the data whose target value are close to 0. The models with the deficient parameter induce this tendency because the errors remain large. The models with the excess parameter can decrease the errors on the model F_1 and F_2 , but a large number of iterations are required to converge. Without enough learning, parameters of model F_1 and F_2 are not consistent with each other because the errors remain large.

We obtain two models after learning, whose structures are the same. If the model well approximates the true structure of the given data, we can expect that the unknown parameters of the model F_1 take nearly the same values as those of F_2 . An unbiasedness criterion with Distorter is defined as follows:

$$UCD = \sum_{j=1}^n (F_1(x_j) - F_2(x_j))^2 \quad (13)$$

where n is the number of the given data.

The UCD has relatively low noise immunity because the approximating properties of all models are approximately identical on the given training data. The quadratic errors are small and their differences are also small for nonlinear models of any structure. The differences of models of any structure will appear on the range of interpolation and extrapolation. This is precisely the situation in which the difference between the model outputs become significant. This motivated the development of αUCD which is widened the interval of summation by introduction of the noise immunity coefficients α , $\alpha = 2 \sim 5$. αUCD is defined as follows:

$$\alpha UCD = \sum_{j=1}^{\alpha \times n} \{F_1(x_j) - F_2(x_j)\}^2 \quad (14)$$

where x_j 's, $j > n$ are the newly generated data which are randomly chosen from unit interval $[0,1]$ and are not used in the training. αUCD is consequently more immune to noise. The model which minimizes UCD and αUCD is selected.

4 Procedure to Operate The System

This system measures the change in the user's head position (i.e., extension/flexion and rotation) and translate this change into a displacement of the cursor on the computer screen. The proposed system enables the sensitivity tuning of cursor and drawing document consistently, by clicking the button on the MENU Window shown in Fig.5. By clicking "3 POINTS TUNING", "3 POINTS", "1 POINT TUNING" and "1 POINT" buttons, the sensitivity tuning for tracking 3 points, the input with sensitivity

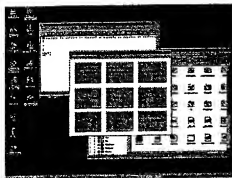


Figure 5: Windows screen and "MENU" Window

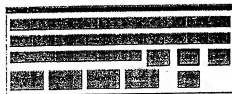


Figure 6: The on-screen keyboard

tuned by 3 POINTS TUNING, the sensitivity tuning for tracking 1 point and the input with sensitivity tuned by 1 POINT TUNING can be implemented, respectively. Furthermore, the on-screen keyboard shown in Fig.6 can be used by clicking "KEYBOARD" button.

The on-screen keyboard is one of input devices which input the characters by clicking the buttons on the screen, instead of pushing the key-tops on the conventional keyboard. By using the on-screen keyboard of proposed system, we can input Japanese characters, uppercase letters, lowercase letters and symbols. By using it together with the editor software, we can draw and print documents.

The procedure to tune the cursor sensitivity by tracking 3 points is as follows: 3 different colored markers are pasted on the user's head or cap as shown in Fig.7. Setting the CCD camera so that the photographic subject does not disappear from the screen. The seated user looks toward CRT, and clicks the "3 POINTS TUNING" button on the MENU Window. (The tracking of the coordinates of markers be-



Figure 7: The locations of markers

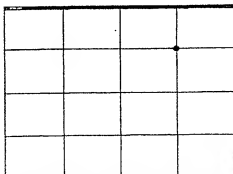


Figure 8: The window for sensitivity tuning

gins at the same time.) When "START" button is clicked, the target red point appears as shown Fig.8. The user turns the face to the point and presses "F1" key. Thus we have a input/output pair of training data, which consists of the coordinates of 3 markers and the coordinates on the CRT when the "F1" key is pressed. This procedure is repeated for 25 points which are placed uniformly on the CRT. The learning of NF-GMDH starts automatically.

We assume that the coordinates on the CRT are normalized in unit square $[0,1]^2$ so that the upper left and lower right of the CRT correspond to (0,1) and (1,1), respectively. The target of red point appears randomly on the lattice points which divides the screen into 25 rectangles of equal size, and their coordinates (X,Y) are represented as $x^* = (x^*, y^*)$. Let the coordinates of 3 markers be $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, $z_3 = (x_3, y_3)$, which are the input vector of NF-GMDH.

Two NF-GMDH models with six input vari-

Table 1: α UCD for the subject A

Layers	x axis	y axis
1	0.300	0.304
2	0.062	0.050
3	0.080	0.193
4	0.124	0.200

Table 2: α UCD for the subject B

Layers	x axis	y axis
1	0.179	0.304
2	0.144	0.062
3	0.211	0.072
4	0.174	0.114

Table 3: α UCD for the subject C

Layers	x axis	y axis
1	0.276	0.045
2	0.165	0.036
3	0.177	0.037
4	0.213	0.056

ables are used to calculate the coordinates (X,Y). The input vector for each network is $x = (x_1, y_1, x_2, y_2, x_3, y_3)$ and desired output are x^*, y^* , respectively. After 5000 learning iterations, the model which minimizes α UCD is selected.

5 Selection of Optimum Number of Layers in NF-GMDH

Three male subjects, whose ages ranged from 23 to 26 were used in this experiment. All were considered novices in the use of head-controlled computer input devices. One of them has a disability on his right upper extremity. 100 input/output pairs of data were obtained through the developed system for sensitivity tuning shown in Fig.8. Tables 1 ~ 3 show the α UCD after 5000 learning iterations. The number of layers which minimizes α UCD is 2 for all subject. Thus the 2 layered NF-GMDH is used for the cursor sensitivity tuning.

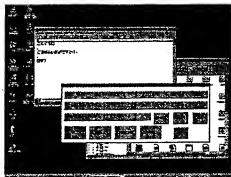


Figure 9: Windows screen in which the on-screen keyboard can be used

Table 4: Comparison of times to input the sentence (sec.)

tuning with	3 points	1 point
subject A	124	168
subject B	155	210
subject C	137	177
average	138	184

6 Evaluation by Using On-screen Keyboard

An input sentence composed of 21 Japanese characters, which means "the study of information processing using neural networks" is used for performance test. The measured time includes the correction time when the subject input incorrect characters.

This experiment is performed by the sensitivity tuning with 3 points and 1 point, respectively. In the sensitivity tuning with 1 point [4], the marker is placed on the top of the subject's nose and the sensitivity tuning is implemented with 2 RBF network in which 9 Gaussian bases are used in each network, which corresponds to the conventional ones mentioned in Section 1.

The subject's seated posture is adjusted toward the CRT. And the CCD camera is set aslant on the right front of the subject.

Table 4 shows the result. The time to input the sentence by 1 point method is longer than that by 3 points method. This is because the

sensitivity of the cursor movement changes depending on the subject's movement direction. Though the subject B has a disability on his right upper extremity, he could input the sentence as fast as others.

7 Conclusion

We have developed a pointing device that translates handicapped person's movements into direct movements of the computer's cursor by measuring several color markers on the user's head. The pointing device which we developed provides the handicapped with convenient coordinate transformation and customised cursor sensitivity by the neurofuzzy GMDH. The equipped CCD camera can be set slantingly. Further developments to simplify the sensitivity tuning and the three dimensional cursor movement are currently proceeding.

References

- [1] Brown, M. and Harris, C., "Neurofuzzy Adaptive Modelling and Control", Prentice Hall, New York, 1994.
- [2] Ergezinger, S. and Thomsen, E., "An Accelerated Learning Algorithm for Multi-layer Perceptrons: Optimization Layer by Layer", IEEE Trans. on Neural Networks, Vol. 6, No. 1, 1995, pp. 31-42.
- [3] Farlow, J., "Self-Organizing Methods in Modeling -GMDH Type Algorithms-", Marcel Dekker, New York, 1984.
- [4] Ichihashi, H., Inoue, J., Ohtani, T. and Miyoshi, T., "Neurofuzzy Mouse with Measurement of Head Positions", Proc. of 12th Fuzzy System Symposium, 1996, pp. 258-260, in Japanese.
- [5] Ichihashi, H. and Turksen, I. B., "A Neuro-Fuzzy Approach to Data Analysis of Pairwise Comparisons", Int. J. of Approximate Reasoning, Vol.9, No. 3, 1993, pp.227-248.

- [6] Ivakhnenko, A. G., "Polynomial Theory of Complex Systems", IEEE Trans. on Syst. Man Cybern., Vol. SMC-1, No. 4, 1971, pp. 364-378.
- [7] Ivakhnenko, A. G., Vysotskiy, V. N. and Ivakhnenko, N. A., "Principal Versions of the Minimum Bias Criterion for a Model and an Investigation of Their Noise Immunity", Soviet Automatic Control, Vol. 11, No. 1, 1978, pp. 27-45.
- [8] Ivakhnenko, A. G.: "Heuristic Self-Organization Systems in Engineering Cybernetics," Tekhnika Press, Kiev, 1971, in Russian.
- [9] Ivakhnenko, A. G.: "The Group Method of Data Handling, A Rival of the Method of Stochastic Approximation," Soviet Automatic Control, Vol. 1, 1968, pp. 43-55.
- [10] Moody, J. and Darken, C. J., "Learning with Localized Receptive Fields", Proc. 1988 Connectionist Models Summer School, San Mateo, CA., 1988.
- [11] Moody, J. and Darken, C. J., "Fast Learning in Networks of Locally-Tuned Processing Unit", Neural Computation, Vol. 1, 1989, pp. 281-294.
- [12] Mucciardi, A. N., "Neuromine Nets as the Basis for the Predictive Component of Robot Brains", Cybernetics, Artificial Intelligence and Ecology, The Macmillan Press LTD., London, 1972, pp. 159-194.
- [13] Ohtani, T., Ichihashi, H., Nagasaka, K. and Miyoshi, T., "Successive Projection Method for Fast Learning Algorithm of Neurofuzzy GMDH", Proc. of the 4th. Int. Conf. on Soft Computing, 1996, pp.432-435.
- [14] Ohtani, T., Ichihashi, H., Nagasaka, K. and Miyoshi, T., "Successive Projection Method for Learning of Neurofuzzy GMDH", J. of Japan Society for Fuzzy Theory and Systems, Vol.9, No.4, 1997, pp. 472-484, in Japanese.
- [15] Ohtani, T., Ichihashi, H., Nagasaka, K. and Miyoshi, T., "Function Approximation by Neurofuzzy GMDH with Error Backpropagation Learning -Empirical Comparisons of Approximation Accuracy with Multilayered Neural Networks-", J. of Japan Industrial Management Association, Vol.47, No.6, 1997, pp. 384-392, in Japanese.
- [16] Ohtani, T., Ichihashi, H., Miyoshi, T. and Nagasaka, K., "Selection of Optimum Number of Layers in Neurofuzzy GMDH with Distorter", Proc. of the 14th Int. Conf. on Production Research, Vol. 2, 1997, pp. 1390-1393.
- [17] Poggio, T. and Girosi, F., "Regularization Algorithms for Learning That Are Equivalent to Multilayer Networks", Sciences, Vol. 274, 1990, pp. 978-982.
- [18] Sugeno, M. and Kang, G. T.: "Structure Identification of Fuzzy Models," Fuzzy Sets and Systems, Vol. 28, No. 1, 1988, pp. 15-33.

Real-Time Facial-Feature Tracking Based on Matching Techniques and Its Applications

Hiroshi Sako, Mark Whitehouse, Anthony Smith, and Alistair Sutherland

Hitachi Europe Ltd., Research & Development Centre, Dublin Laboratory (HDL)

O'Reilly Institute, Trinity College, Dublin 2, IRELAND

Tel: +353-1-6798911, Fax: +353-1-6798926, E-mail: hiroshi@hdl.ie

Abstract

This paper describes a method of real-time facial-feature extraction which is based on matching techniques. The method is composed of facial-area extraction and mouth-area extraction using colour histogram matching, and eye-area extraction using template matching. By the combination of these methods, we can realize real-time processing, user-independent recognition and tolerance to changes of the environment. Also, this paper touches on neural networks which can extract characteristics for recognizing the shape of facial parts. The methods were implemented in an experimental image processing system, and we discuss the cases that the system is applied to man-machine interface using facial gesture and to sign language translation.

1: Introduction

Facial image recognition is an important technology in person identification for security systems, facial gesture understanding for advanced man-machine interface, and image coding for facial image data transmission. In general, there are two types of approaches in facial image recognition, synthetic facial image recognition and analytic facial image recognition. The former approach regards a whole facial image as one pattern and the image is, for example, classified by neural networks[1][2] or discriminated by feature components extracted by KL expansion[3][4]. An advantage of this approach is its simplicity. A disadvantage is its homogeneity in the structure which makes it harder to change the process flexibly. The latter approach includes the conventional bottom-up approach[5][6] which is generally composed of three steps: segmentation of the facial parts, shape analysis of the parts and statistical decision. An advantage of this approach is its analyzability by which you can modify the method flexibly while a disadvantage is its dependency on some particular problem. We prefer the latter approach, shown in figure 1, because it requires at least three sub-technologies corresponding to each step and they may be applied to other domain applications.

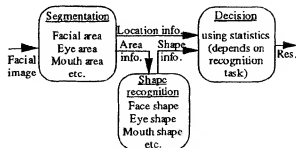


Figure 1. Facial image recognition.

In this paper, we mainly focus on a method[7] of facial-feature extraction in facial gesture understanding, which corresponds to the first step of the bottom-up approach, and also touch on a method[8] of shape recognition of the facial part, which corresponds to the second step of the bottom-up approach. There are several types of facial-feature extraction methods including template matching[9], neural networks[10], colour-based approach[11], and deformable templates approach[12][13]. When we consider real systems which use facial gesture understanding (for example, applications to man-machine interface), real-time processing, user-independent recognition and tolerance of the change of atmosphere according to time are indispensable. Our approaches to these research issues are: (1) To develop computationally inexpensive methods based on matching and to combine them effectively for real-time processing, (2) To use multi-models and templates for user-independent recognition, (3) To use the information in the previous frame to ensure tolerance to changes in environment such as illumination change with respect to time.

At the end of this paper, we discuss actual applications of our developed techniques to the field of advanced man-machine interface.

2: Facial-feature extraction

The overall process in the facial-feature extraction method is shown in figure 2, which is composed of the facial-area extraction and mouth-area extraction using

colour histogram matching (CHM), and eye-area extraction using template matching (TM). In the case of every frame except the first, the model and the template in CHM and TM are occasionally renewed according to feature extraction results of the previous frame. We call this process sequential matching (SM). As for the first frame image, we prepare multi models for CHM and multi templates for TM to achieve user independence. Table 1 shows the relationship between CHM, TM and SM.

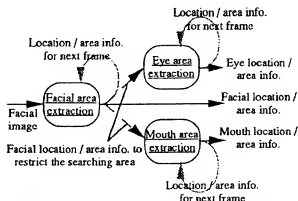


Figure 2. Overall process in the facial-feature extraction method.

Table 1. Facial-feature extraction methods

Facial feature	Facial area	mouth	eyes
Information used for extraction	Colour	Colour	Pixel density distribution
Extraction Methods	Colour Histogram Matching (CHM)	Template Matching (TM)	
Model & Temp. acquisition	Sequential Matching (SM)		
Extraction rate	16/16(100%)	15/16(94%)	(Under estim.)

Using the detected information about the facial location and area, we can set reasonable limited search areas for eye and mouth extraction, which has great effect on recognition accuracy and processing complexity. The facial-area and the mouth-area extraction part actually include a normalization of the colour image, and the eye-area extraction part has a translation from the colour image to the grey image as a pre-processing, though figure 2 doesn't describe them because they aren't topics in this paper.

2.1: Facial-area and mouth-area extraction

As shown in table 1, we use colour information to extract the position of the face and the mouth. In other words, we search for specified coloured parts as objects to be extracted. The reasons why we use colour information are: (1) the facial skin and the mouth colour distribution are almost unique in one human race unless the user wears makeup, (2) the shape of the object such as the mouth and the face is easily changed by its movement and it's not

easy to use shape information for extraction, (3) the search for a specific colour is realized by simple operations and good for real-time processing.

To search for specified coloured pixels, we use the colour histogram matching (CHM) method based on the colour indexing technique[14], which is shown in figure 3. This method is based on the probability that each examined pixel p in the input facial image may belong to the pre-determined model image by calculating the following ratio $R(j)$.

$$R(j) = I(j) / M(j) \quad \text{if } M(j) \geq I(j),$$

$$R(j) = M(j) / I(j) \quad \text{if } M(j) < I(j),$$

where j is the bucket number which is indexed by the RGB value of the pixel p , $I(j)$ is the value at the bucket j of the input image RGB-histogram and $M(j)$ is the value at the bucket j of the template image RGB-histogram. If the model image is a unique subset of the input image in respect of colour distribution, then the ratio $R(j)$ becomes 1.0 when the pixel p is in the model image, but otherwise the ratio $R(j)$ becomes 0.0. Therefore, if only you make a back-projected image P by calculating the ratio $R(j)$ at each pixel of the input image, you can basically extract the area with the maximum $R(j)$ ($=1.0$) corresponding to the model image.

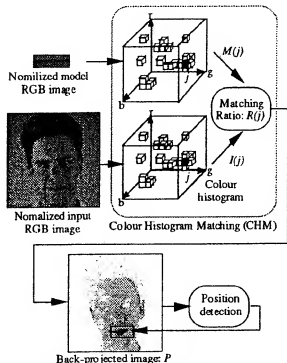


Figure 3. Mouth-area extraction.

Figure 3 shows an outline of the mouth-area extraction process which includes CHM and a position detection part. The position detection of the mouth is composed of two

stages: the coarse and fine position detection stages. The coarse position detection is realized by detecting the maximum point of the image P' which is the smoothed image of P by local-averaging. The size of the kernel of the local-averaging is almost the same as the size of the model. At the fine detection stage, a limited area of the image P the centre of which is located at the maximum point is thresholded into a binary image and it is projected onto the x and y directions. By detecting the edge of each projection, we can detect the detailed mouth boundary.

To reduce dependence on user's mouth colour in the processing, we prepare several typical mouth models in the form of colour histograms $M_k(j)$ ($k=1, N, N'$: number of models). In this case, we calculate each ratio $R_k(j)$ according to each model and select the ratio $R_k(j)$ with maximum value as $R(j)$ at every j .

One of the advantages of this method is that each step in the method is computationally inexpensive and can be completed within a frame interval by using popular image processing equipment. Therefore it is possible to realize pipeline processing at frame-rate.

We can realize facial-area extraction by a similar process using skin-colour models.

2.2: Eye-area extraction

As shown in Table 1, we use shape information (pixel density distribution in image space) to extract the position of the eyes. In another words, we search for a part with a specific pixel density distribution as the object to be extracted. The reasons why we use shape information are; (1) the eye shape is almost unique in the face area, even if the user closes his eyes, (2) the search for a part with a specific shape is realized by the following simple operation, the calculation of which can be easily completed in a frame interval, and is good for real-time processing.

$$D(x, y) = \sum_i \sum_j \{G(x+i, y+j) \cdot T(i, j)\}^2,$$

where $D(x, y)$ expresses the degree of matching between the input gray image $G(x+i, y+j)$ and the template pattern $T(i, j)$ at the location of (x, y) in the input image, and (i, j) are coordinates of the template. Figure 4 shows the eye extraction process which includes TM and the following position detection by searching for the minimum point of $D(x, y)$ after some noise reduction.

To reduce dependence on user's eye shape in the processing, we prepare the several typical eye templates in the form of image $T_k(i, j)$ ($k=1, N, N'$: number of templates). In this case, we calculate each degree $D_k(x, y)$ corresponding to each template and select the degree $D_k(x, y)$ with maximum value as $D(x, y)$ at every (x, y) .

Figure 5 shows the facial-feature extraction results of different people. In this experiment, we processed the first frame image (512*512 RGB pixels) by CHM and TM with no information about the previous frame. As shown

in Table 1, the recognition rate is above 94 % though the number of the examined facial images is only 16.

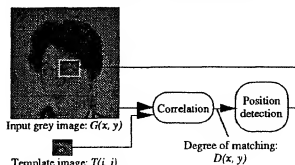


Figure 4. Eye-area extraction.

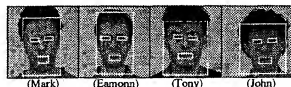
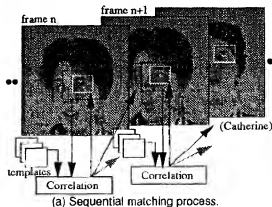
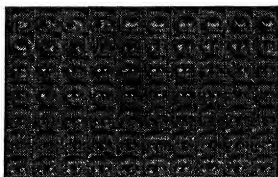


Figure 5. Facial-feature extraction results.

2.3: Sequential matching by the renewal of models and templates

Figure 6(a) shows the sequential matching (SM) process in which one of templates is renewed following eye-extraction. Once you can detect the eye in the previous frame correctly, then the template matching in the present frame is done in a restricted area about the previous eye position by templates including a template derived from the previous frame. The reason why we renew the template is to ensure tolerance to gradual change of environment such as the change of the illumination. Figure 6(b) shows the extraction results of 70 frames which corresponds to 2.3 seconds.





(b) Detected and stored sequential eye-parts.

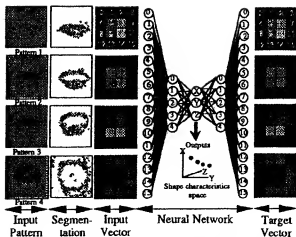
Figure 6. Sequential matching.

In the SM of mouth extraction, we renew one of histogram models of the mouth. That is obtained from the detected mouth image in the previous frame.

3: Facial part shape recognition

Quantitative expression of the shape of the facial part is required to recognize facial gesture expression. For example, lip reading[15], which may be used in an advanced man-machine interface, needs to analyze the time series data expressing the mouth shape quantitatively. Our approach to measuring the shape is to use hour-glass type neural networks[16]. The reasons why we use these are: (1) it is difficult to determine specific characteristics suitable for shape recognition, and the characteristics usually depend on the kind of objective image. Neural networks have a potential to be applied to small domain task such as shape measurement of the facial part with few heuristics. (2) Real-time processing is possible when we prepare dedicated hardware for the neural networks.

Figure 7(a) shows the learning and the recall process. This neural network has 5 layers and $16 \times 5 \times 3 \times 5 \times 16$ neurons in total. We prepare several typical images in the learning stage. Each image is segmented into the mouth-area (the back-projected image) using colour analysis (CHM) and the segmented image is divided into 16 sections, the average values of which are the elements of the input vector for the neural network. The target vector is same as the input vector. At the end of this stage, the outputs of the neurons in the third layer, X, Y, Z , are regarded as shape characteristics of the corresponding input image. In the recall stage, the facial-feature image, for example, the mouth part image, which is detected at the facial-feature extraction stage described in section 2, is input to the network and its shape is judged by the outputs of the third layer X, Y, Z . The ellipses in figure 7(b), which are generated by the detected shape characteristics, indicate the results of shape recognition.



(a) Learning and recall process of neural networks.



(b) Results of shape recognition. (Hiroshi)

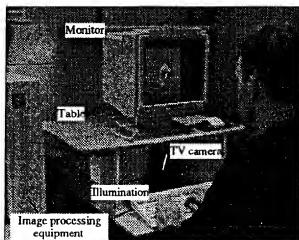
Figure 7. Neural networks for shape recognition.

4: Applications of real-time facial-feature extraction

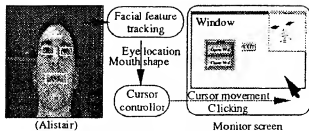
There are many applications of the real-time facial image processing techniques which we have developed, including security systems, advanced man-machine interface, facial image coding for data transmission and so on. We are now interested in an advanced man-machine interface using facial gestures, and also interested in a sign language translation system to allow communication between deaf and non-deaf persons. The system sometimes requires to know which facial part the finger points to. There are many expressions which use positional relationships between fingers and facial parts (for example, 'see' 'look' 'taste' 'sneeze' etc. in British Sign Language).

Figure 8(a) shows our experimental system which is composed of a TV camera, a monitor, some image processing equipment (DataCube) and a special table with illumination. This system can track the facial-area, eye and mouth-area, and the coloured fingertips of the user in real-time (about 20 frames / second).

Figure 8(b) illustrates the case that the system is applied to act to replace the mouse. Mouse operations such as cursor movement and clicking operation are controlled by eye position and mouth shape change, respectively. Positional relationships required in a sign language translation system can easily realized by the geometrical calculation of the detected positions of facial parts and coloured fingertips. (These applications will be shown in the video in detail.)



(a) Real-time facial-feature tracking system.



(b) Cursor control by facial gesture.

Figure 8. An application of the real-time facial-feature tracking system.

5: Conclusions

We have developed a real-time facial-feature extraction method based on matching techniques. The method is composed of facial-area extraction and mouth-area extraction using colour histogram matching, and eye-area extraction using template matching. These methods are suitable for real-time processing because of their computationally inexpensiveness. We used multi-models and templates to reduce user dependency. To tolerate changes in the environment over time (such as change in illumination), we actively used information about the previous frame to modify the model and template. Moreover, in this paper, we touched on shape recognition using the hour-glass neural networks.

The methods were implemented on an experimental hardware system. The system has been applied to the man-machine interface using facial gesture and to a sign language translation system to recognize positional relationships between user's fingers and facial parts.

Further research includes real-time facial expression recognition by analyzing time series data, and its application to real-time synthesis of facial expression in computer graphics.

Acknowledgements

The authors wish to thank Dr. M. Sugie of the Central Research Laboratory, Hitachi, Ltd., Dr. Y. Kuwahara and Dr. M. Abe of Hitachi Europe Ltd. for their encouragement and support throughout this research, and would also like to thank Dr. M. Ejiri of the Central Research Laboratory, Hitachi, Ltd. for his helpful suggestions, Mr. S. Weavers and Mr. J. Ford for their support in the implementation of the DataCube.

References

- [1] M. K. Fleming and G. W. Cottrell: "Categorization of faces using unsupervised feature extraction," *Proc. of IJCNN90*, II, pp. 65-70 (1990).
- [2] M. Kogusi: "Human-Face identification by neural network," *Proc. of Japanese national conference of IEICE*, D-407 (1990). (in Japanese).
- [3] M. Kirby and L. Sirovich: "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Trans. on PAMI*, Vol.12, No.1, pp. 103-108 (1990).
- [4] M. A. Turk and A. P. Pentland: "Face recognition using eigenfaces," *Proc. of CVPR91*, pp.586-591 (1991).
- [5] Y. Kaya and K. Kobayashi: "A basic study on human face recognition," in *Frontiers of Pattern Recognition* (S. Watanabe, Ed.), Academic Press, pp. 263-289 (1971).
- [6] T. Sakai et al.: "Computer analysis and classification of photographs of human faces," *Proc. of First USA-Japan Computer Conference*, pp. 66-62 (1972).
- [7] A. Sutherland, A. Smith and H. Sako: "Real-time eye-tracking using pipeline architecture," 4th International conference on visual search, Eindhoven, The Netherlands, Aug. (1994). (to be submitted).
- [8] A. Smith, M. Whitehouse and H. Sako: "Mouth shape recognition using hour-glass neural networks," 4th International conference on visual search, Eindhoven, The Netherlands, Aug. (1994). (to be submitted).
- [9] R. Brunelli and T. Poggio: "Face recognition: features versus templates," *IEEE Trans. on PAMI*, Vol.15, No.10, pp. 1042-1052 (1993).
- [10] J. M. Vincent, J. B. Walts and D. J. Myers: "Location of features in images using neural networks," *BT Technol. J.*, Vol.10, No.3, pp. 7-15 (1992).
- [11] T. Sasaki, S. Akanatsu and Y. Suenaga: "Face image normalization based on color information for automatic face recognition," *Proc. of Japanese national conference of IEICE*, D-190 (1991). (in Japanese).
- [12] A. L. Yuille: "Deformable templates for face recognition," *Journal of cognitive neuroscience*, Vol.3, No.1, pp. 59-70 (1991).
- [13] I. Craw, D. Toek and A. Bennet: "Finding face features," Technical Report 92-15, Departments of mathematical sciences, Univ. of Aberdeen, Scotland (1991).
- [14] M. J. Swan and D. H. Ballard: "Color Indexing," *International Journal of Computer Vision*, Vol.7, No.1, pp. 11-32 (1991).
- [15] A. P. Pentland and K. Mase: "Lip reading: Automatic visual recognition of spoken words," *Proc. of Image Understanding and Machine Vision*, Optical Society of America, PP. 12-14 (1989).
- [16] G. W. Cottrell and P. Murru: "Principal component analysis of images via back propagation," *SPE, 1001, Visual communications and image processing '88*, pp.1070-1076 (1988).



Computer Interface to Use Head and Eyeball Movement for Handicapped People

* Osamu TAKAMI,

** Kazuaki MORIMOTO,

*** Tsumoru OCHIAI and ** Takakazu ISHIMATSU

* Technology Center of Nagasaki Prefecture, Ohmura 856 JAPAN

** Dept. Mechanical Systems Engineering, Nagasaki University, Nagasaki 852 JAPAN

*** Ube Technical College, Ube, Yamaguchi 755 JAPAN

ABSTRACT

In this paper we propose one computer interface device for handicapped people. Input signals of the interface device are movements of eyeballs and head of the handicapped. The movements of the eyeballs and head are detected by an image processing system. One feature of our system is that the operator is not obliged to wear any burdensome device like glasses and a helmet. The sensing performance of the image processing of the eyeballs and head is evaluated through experiments. Experimental results reveal the applicability of our system.

1. INTRODUCTION

Nowadays, the increase of the aged people has become one of the notable social problems, and it should be noticed that many of them are handicapped in some meanings. Therefore, development of supporting devices and care equipments for the handicapped is desired. One difficult problem related with the development of such supporting devices is that the every handicapped people has various physical abilities. Therefore, supporting devices should be developed considering the ability of every handicapped people. One possible solution to cope with this problem is to develop interface devices between the handicapped people and the computer. Once it becomes possible for the handicapped people to communicate with the computer using some interface device friendly, the computer becomes a big help for them. Already some interface devices are developed for the handicapped to communicate with the computer using some remaining physical abilities. Some devices use the simple breathing actions or patting actions of the handicapped as input signals. For the heavy handicapped people one interface device to use eyeball movements as input signals is proposed, considering that the eyeball movements are

relatively easy to control even for the heavy handicapped people. Several techniques to measure the eyeball movements [1,2,3] are also reported. One technique is to use optical sensors those are mounted on a specialized helmet or glasses. However, wearing a helmet and glasses are not preferable and sometimes unacceptable for the handicapped because of some physical reason.

In this paper we propose an interface device to use eyeball movements as input signals. One feature of our system is that the image processing technique is employed to detect eyeball movements. Therefore, the camera is settled apart from the operator without forcing the operator to wear special glasses or a helmet. The camera is used to detect the orientation of the eyeball and the blinking of the eyes. The information about eyes obtained by the image processor are used as input signals of the environmental control system for the handicapped people. This environmental control system can control the switching of the electrical devices like TV, radio, lights and so on. We named this system "Eye-Controller". Experiments to evaluate the accuracy of the detection of the eyeballs revealed enough performance to use for the handicapped people. Another feature of our system is that the 3-dimensional movements of the head is also used as input signals to the interface device. The information about head movement is useful to enhance the accuracy of detecting the eyeball movements and also simplify the input operation to the interface system. In case that the operator is able to move his head intentionally, the head movement simplify the input operation remarkably.

In order to enable stable 3-dimensional measurement of the head movement, two marking points are settled on the forehead of the operator. One camera obtains the image of the operator's face. Of course, both eyes of the operator and two marking points on the forehead need to be included in the image. Computer analyses the images and determines the 3-dimensional posture and position of the head.

We tested our interface device at a training facilities for the

handicapped. Experimental results was satisfactory.

2. CONFIGURATION OF INTERFACE DEVICE

The interface between the handicapped people and the computer using the head and eyeball movements can be realized using an image processing technique. In Fig.1 the configuration of our interface device is shown, where a TV camera is settled in front of the operator to detect the image of the operator. The operator can select the menu which is displayed on the TV monitor. One example of the menu on the TV monitor is shown in Fig.2, where five menu items are shown. A handicapped operator faces the display monitor and selects one menu item among

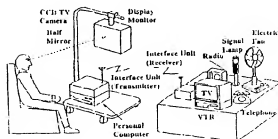


Fig.1 Configuration of interface device

menu items by gazing at it and by blinking his eyes for confirmation. Suppose the operator selects the third item (Telephone) in Fig.2, the next sub-menu is displayed on the display monitor as shown in Fig.3. Suppose the operator selects the second menu item (family), the interface device starts to connect the telephone line with his family automatically. Similarly every item on the menu has sub-menu if necessary.

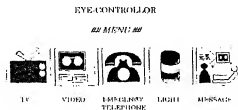


Fig.2 Main menu

In Fig.4 the block diagram of our interface device is shown. In order to enable fast and tactical image processing, the image data are processed using a personal computer (NEC PC 9801 AN, 90Mhz Pentium processor) and a real-time image



Fig.3 Sub Menu (Telephone)

processor. The real-time image processor is composed of a FPGA(Field Programmable Gate Array) and its main function is real-time labeling operation(1/30 sec) on the binary image data(512h × 256v). Due to this function, fast detection of the eyeball movement and marking points on the forehead of the operator are simplified remarkably.

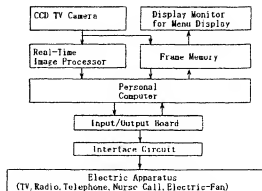


Fig.4 Block diagram

3. MEASUREMENT OF EYEBALL MOVEMENT

3.1 Algorithm to detect eyeball movement

Two important image processings of our interface device are how to detect the position of the eyes in the image and how to detect the orientation of the eyes. Considering that the operator is apt to blink his eyes during the operation of the interface device, one simplified technique is introduced to detect the position of the eyes. At the beginning of the operation, the operator is requested to blink his eyes a couple of times. During the blinking, the real-time image processor performs the subtraction of two-sequential images. For the resultant image, conversion into binary image data and the labeling procedure is executed. Suppose head movements are negligible, position of the eyes can be detected easily using the above labeling procedure.

Once the positions of the eyes in the image are detected, the image of the eyeball is analysed in detail and orientation of the eyeball is measured. It should be noticed that detection of the eye position using eye-blinking needs not to be executed in the subsequent measurements. Because the image processing is executed quickly (1/30 sec) and the position of the eyes at the next measuring instant is easily estimated from the previous image analysis.

In order to detect the orientation of the eyeball, the parameter R defined as $R = A/B$ is introduced, where A is the distance between the beginning of the eye and the center of the pupil and B is between the beginning of the eye and the tail of the eye. This definition is illustrated in Fig 5(a). The positions of the pupils are detected as the midpoint of two cross points, which are explained in Fig 5(b).

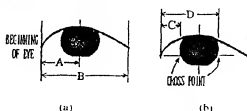


Fig 5 Definition of parameters

3.2 Experiment of parameter R

Relation between the orientation of the eyeball and the parameter R is checked by experiments.

The display monitor was settled 50cm apart from the operator. Five menu items are horizontally arranged with 52mm interval on the display monitor as is illustrated in Fig 6)

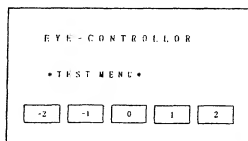
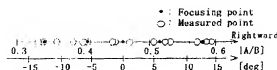


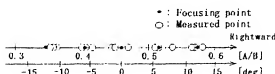
Fig 6 Test menu

Under the condition that an operator gazes at the center of menu items, the parameter R (orientation of eyeball) is measured by the image processing.

In Fig 7(a) experimental results obtained by measuring only right eye are shown. In Fig 7(b) experimental results obtained by measuring both eyes are shown. These results mean that



(a) Focusing angle measured using right eye



(b) Focusing angle measured using both eyes

Fig 7 Focusing points and parameter R

measuring one eye is enough to recognize which menu item is gazed at. The results also mean that the detecting performance increase if both eyes are measured.

Considering the experimental results, handicapped peoples tested our environmental control system. Only in the case that the lead of the handicapped people was fixed enough, the performance was satisfactory.

5. MEASUREMENT OF HEAD MOVEMENT

From the experimental results, it became clear that orientation of the eyeball can be detected satisfactory only in the case that the movement of his head is negligible. However, movements of the head is inevitable. Practical interface device should deal with the movements of the head in order to enlarge the application field of this system. It is also important that if the movement of head can be used as the control signal of the interface device, the device becomes more friendly with the handicapped.

In the followings we explain how to detect the 3-dimensional position and posture of the head. Two marks are attached on the forehead of the operator so that these two marks and two eyeballs compose an parallelogram (see Fig. 8). By measuring the direction of these two marks and two eyeballs using a TV camera, the three-dimensional posture and position of this parallelogram can be readily determined as follows.

Suppose the position vectors of the four corners of the parallelogram are $p_i (i=1,2,3,4)$. The direction vector of these four corners can be measured as $q_i (i=1,2,3,4)$ by the camera. From the geometric relation that p_1-p_2 runs parallel with p_4-p_3 , p_i are determined as the $p_i-k_i q_i$ where k_i are scalar numbers. Scalar numbers k_i can be readily obtained from the following relation and actual length of p_1-p_2 and p_4-p_3 .

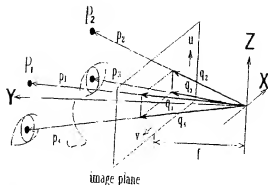


Fig. 8 Geometry of two marking points

$$k1/k4 = (q4 \times q3) \cdot q2 \cdot (q1 \times q3) \cdot q2$$

$$k2/k4 = (q4 \times q3) \cdot q1 \cdot (q2 \times q3) \cdot q1$$

$$k3/k4 = (q4 \times q2) \cdot q1 \cdot (q3 \times q2) \cdot q1$$

Since 3-dimensional positions of the marks on the forehead and eyes can be determined, the 3-dimensional position and posture of the head, and also distortion of the parallelogram are determined. It is important that the distortion of the parallelogram gives the information about orientation of the face.

We had experiments using a mannequin instead of a operator. The mannequin head was settled toward five different directions. Firstly, the head was settled rightward with 15 degree and the orientation was measured by the image processing. Similar experiments were executed with different orientation. The results are shown in Fig. 9. While the results were obtained using a mannequin, the accuracy to detect the orientation of the eyes was less than ± 5 degree.

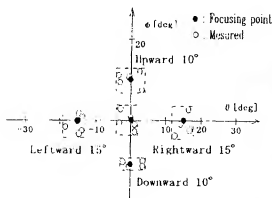


Fig. 9 Experimental result to detect orientation of head

6. FIELD TEST

In Photo 1 one Eye-Controller is shown. Considering the usage at the bedside, a thin liquid crystal display monitor, a TV camera and half mirror are united in a box. Three heavy handicapped people tested the interface device. The first and the second people laid himself on the bed during the experiment (see Photo 2). Using the movements of eyeball, they succeeded to communicate with the computer without any trouble. The third woman on the wheel chair had some difficulty to communicate with the computer (see Photo 3). One reason of this difficulty was that she was not able to gaze at one desired menu item because of convulsive movements. However, she could operate the system using head movements. She moved a cursor to a menu item on the display monitor by facing to the right and left and clicked the desired menu by nodding her head.

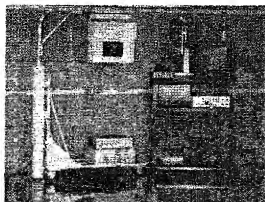


Photo 1 Eye-Controller System



Photo 2 Test scene 1



Photo.3 Test scene 2

7. CONCLUSIONS

A new environmental control system to use eyeball movement without wearing any device was developed. Detection of eyeball movements and the direction of the head became possible by the image processing. Even if an operator's face inclines or moves, it became possible to control personal apparatus by using eyeball movements. Moreover, it became possible to operate the system by using head movements. We built a test system in order to apply our environmental control system for the daily use. It was possible for handicapped people to operate our environmental control system in the bed side. The system needs to be adequately modified considering the physical abilities of the operator. Furthermore, the system needs to be more compact, cheap and reliable. These themes are now under study.

References

- [1] M.Yamada and T.Fukuda,"A Word Processor and Peripheral Controller Using Eye Movement",Trans.ICE (in Japanese),Vol.J69-D, No.7,1986,pp.1103-1104.
- [2] T.Shimonouchi,H.Irie,T.Ishimatsu and O.Takami,"A Robot Control Method by Eyeball Movements",Proc. of A-PVC'93,1993,pp.1038-1041.
- [3] M.Suzuki et al., "Development of An Interface for The Elderly People Using Eye-Motion",Proc. of 3rd Bio-Engineering Symposium (in Japanese), 1994, pp.12-13.
- [4] O.Takami,T.Ishimatsu and T.Shimonouchi," Development of the Environmental Control System by Using Eyeball Movements",Proc. of 1st Asian Control Conference, Vol.3, 1994, pp.415-418.
- [5] H.Aoyama and M.Kawagoe,"A Sensing and Recognizing Method for Directions of Face and Gaze Using Plane Symmetry", Human Interface (in Japanese),Vol.4, No.3, 1989, pp.245-254.